

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
27 December 2002 (27.12.2002)

PCT

(10) International Publication Number  
WO 02/103555 A2

(51) International Patent Classification<sup>7</sup>:

G06F 17/28

(81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(21) International Application Number:

PCT/GB02/02742

(22) International Filing Date:

12 June 2002 (12.06.2002)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

09/883,693

18 June 2001 (18.06.2001) US

(71) Applicant (for all designated States except US):  
PIPEDREAM METASYSTEMS, INC. [US/US];  
1288 Columbus Avenue #153, San Francisco, CA 94133  
(US).

(71) Applicant (for MG only): HOWE, Steven [GB/GB]; 36  
Longacre, Chelmsford, Essex CM1 3BJ (GB).

(72) Inventor; and

(75) Inventor/Applicant (for US only): MANSON, Keith, S.  
[US/US]; P.O. Box 8406, Berkeley, CA 94707 (US).

(74) Agent: HOWE, Steven; Lloyd Wise, Tregear & CO.,  
Commonwealth House, 1-19 New Oxford Street, London  
WC1A 1LW (GB).

(84) Designated States (regional): ARIPO patent (GH, GM,  
KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW),  
Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),  
European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR,  
GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent  
(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR,  
NE, SN, TD, TG).

Published:

— without international search report and to be republished  
upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/103555 A2

(54) Title: COMPUTER SYSTEM WITH NATURAL LANGUAGE TO MACHINE LANGUAGE TRANSLATOR

(57) Abstract: Presented is a system and method for converting or translating expressions in a natural language such as English into machine executable expression in a formal language. This translation enables a transformation from the syntactic structures of a natural language into effective algebraic forms for further exact processing. The invention utilizes algorithms employing a reduction of sequences of terms defined over an extensible lexicon into formal syntactic and semantic structures. This term reduction incorporates both syntactic type and semantic context to achieve an effective formal representation and interpretation of the meaning conveyed by any natural language expression.

## COMPUTER SYSTEM WITH NATURAL LANGUAGE TO MACHINE LANGUAGE TRANSLATOR

### *BACKGROUND OF THE INVENTION:*

The present invention is directed to a system which translates natural (human) language into an abstract formal language. This formal language is explicitly designed to serve as a universal template for further translations into a comprehensive variety of machine languages which are executable in specific operational environments. Extensive efforts have been made, many articles have been published, and many patents have been issued, all directed toward the goal of providing computers with the capacity to understand natural (human) language sufficiently well to respond reliably and accurately to directives issued from human users. Many companies and research groups, such as AT&T, IBM, and Microsoft, and an assortment of academic institutions, are presently working on natural language processing (NLP).

To date, many different approaches have been tried to provide a system which effectively converts natural language to a formal language for computer applications. One such approach is disclosed in an article published by Microsoft Corporation titled "Microsoft Research: Natural Language Processing Hits High Gear" dated May 3, 2000. The article discloses that Microsoft is heavily focused on a database of logical forms, called MindNet (TM), and the creation of a machine translation application. It is stated that MindNet is an initiative in an area of research called "example-based processing", whereby a computer processes input based on something it has encountered before. The MindNet database is created by storing and weighting the semantic graphs produced during the analysis of a document or collection of documents. The system uses this database to find links in meaning between words within a single language or across languages. These stored relationships among words give the system a basis for "understanding", thereby allowing the system to respond to natural language input. MindNet apparently contains the contents of several dictionaries and an encyclopedia to increase its level of understanding. Another approach is disclosed in Microsoft U.S. Patent No. 5,966,686. This approach provides a rule-based computer system for semantically analyzing natural language sentences. The system first transforms an input sentence into a syntactic parse tree. Semantic analysis then applies three sets of semantic rules to create an initial logical form graph from this tree. Additional rules provide semantically meaningful labels to create additional logical form graph models and to unify redundant elements. The final logical form graph represents the semantic analysis of the input sentence.

Yet another, and apparently more common, approach is provided by U.S. Patent No. 5,895,466, wherein a database stores a plurality of answers which are indexed to natural language keys. The natural language device receives a natural language question over the network from a remote device and the question is analyzed using a natural language understanding system. Based on this analysis, the database is then queried and an answer is provided to the remote device.

Applicant is aware that various other approaches toward providing a conversion from natural language to some machine language have been tried. However, the prior art has not provided a truly effective conversion system of this sort.

**SUMMARY OF THE INVENTION:**

Presented is a system and method for converting or translating expressions in a natural language such as English into machine executable expressions in a formal language. This translation enables a transformation from the syntactic structures of a natural language into effective algebraic forms for further exact processing. The invention utilizes algorithms employing a reduction of sequences of terms defined over an extensible lexicon into formal syntactic and semantic structures. This term reduction incorporates both syntactic type and semantic context to achieve an effective formal representation and interpretation of the meaning conveyed by any natural language expression.

The foregoing features and advantages of the present invention will be apparent from the following more particular description of the invention. The accompanying drawings, listed herein below, are useful in explaining the invention.

**BRIEF DESCRIPTION OF THE DRAWINGS:**

FIG. 1 shows the hardware architecture of a computer system comprising the natural language converter of the present invention;  
FIG. 2 shows the general process and data flow of the inventive system;  
FIG. 3 shows a more detailed flow diagram for the inventive system;  
FIG. 4a shows the results of virtual type assignment applied to a sample text;  
FIG. 4b shows the results of actual type assignment for the same text;  
FIG. 5 shows the term reduction sequence for a sample text;  
FIG. 6a shows the sequence of dependency chains for a sample text;  
FIG. 6b shows the associated syntactic tree for the same text; and  
FIG. 7a shows the schema of structures and maps involved in the external interpretation of a text;  
FIG. 7b shows this external interpretation schema as controlled by a metasemantic protocol.

**BRIEF DESCRIPTION OF THE INVENTION:**

Refer to FIG. 1 for an overview of the system architecture. As mentioned above, the inventive system, called METASCIPT (TM), provides a method for translating expressions in a natural language such as English into machine executable expressions. In the embodiment of the system and method to be described, the user inputs text in a natural language through some input device to a known computer system which may comprise a standalone computer system, a local network of computing devices, or a global network such as the Internet, using wired land lines, wireless communication, or some combination thereof, etc. This computer system includes memory for storing data, and a data processor. The text may be entered into the client device or local VDM (Video Display Monitor) (1.1) by any suitable means, such as direct input via keyboard (1.1.1), voice input via speech recognition means (an SR system) (1.1.2), or indirect input via optical scanning (an OCR system) (1.1.3). The natural language text input to the system is passed along the network or local bus (1.3) to a server or local CPU (Central Processing Unit) (1.2) where it is processed in accordance with the inventive method and system. This processed output of the system is then provided to the system for distribution to the original input device (1.1), or to other collateral devices (1.4) which may be one or more digital computers, mobile devices, etc. The inventive system thus comprises a natural language interface to any sufficiently capable digital environment.

Refer now to FIG. 2 for an overview of the process and data flow of the inventive system. The invention will be subsequently discussed in more detail herein below. Natural language text input is entered by the user (2.0) into the internal system (2.1) by means of a text processing module (2.1.1) which parses the text. The output of the text processing module comprises a parsed sequence of preexpressions which is entered into the syntactic processing module (2.1.2) which provides syntactic type information, establishes proper syntactic dependencies between terms in expressions, and represents these expressions as complexes in a syntactic algebra. The output of the syntactic processing module, comprising a sequence of these syntactic complexes, is entered into the semantic processing module (2.1.3) in order to achieve a semantic interpretation of the input text. The output of the semantic processing module, comprising a formal interpretation of the input text, is entered into an external system by means of the external processing module (2.2.1), which finally provides a sequence of executable expressions derived from the input text for use in a specific operational environment (2.2.2).

As noted above, the means for providing text input to the system, such as through a keyboard, scanner, or speech recognition system, are well known in the art and are commercially available. Another standard component of the present system is a text parser, construed here in an extremely narrow sense as a limited process restricted to partitioning text strings into syntactic subcomponents such as paragraphs, sentences, and words. As such, the text parser discussed herein does not provide further linguistic information such as grammatical types, syntactic dependencies, semantic import, etc. Such limited text parsers are standard components of any natural language processing system, and exceedingly well known in the art. Yet another component in the present system which plays a relatively standard role is the lexicon or "electronic dictionary". In general, lexicons are also well known in the art and are discussed in many patents including U.S. Patent Nos. 5,371,807; 5,724,594; 5,794,050; and 5,966,686. However, the notion and function of "virtual" types, which play a significant syntactic categorization role in the passive specification of lexical terms, and hence strongly contribute to the definition of the particular lexicon used in the inventive system, are not standard, and so require careful description. On the other hand, since text input devices, text parsers, and their operation are so well known, they will not be further described in detail herein.

Refer now to FIG. 3, which shows more details of the inventive system. The components, modules, and submodules of the inventive system are enumerated for convenient reference so that the operation and application of the system and method may be described in detail.

As mentioned above, natural language text is entered by the user (3.0) into the text input submodule (3.1.1) of the text processing module (3.1) via any suitable means including a keyboard or a speech recognition system. For the purposes of this discussion, the user input signal is simply some linguistic data stream which is digitized into a string of ASCII characters. This ASCII string is the input text.

In order to clarify the following discussion, it is helpful to note that any natural language text is typically organized into a sequence of paragraphs, each of which is a sequence of sentences, each of which is a sequence of words, each of which is a sequence of characters (alphanumeric symbols). All of this nested syntactic structure must be taken into account if an effective interpretive analysis is to be achieved. The role of the text parser is to determine and then present these nested sequential structures to the system for further processing. Thus in general, the adequate output of the text parser is a sequence of sequences of sequences of ASCII characters. This level of generality, however, tends to obscure the basic points of any useful description of the inventive system, so a technical compromise is adopted herein, whereby any text is considered to comprise a sequence of sentences, or more properly, of expressions, each of which comprises a sequence of words. Until recognized by the system as a meaningful unit of linguistic analysis, however, any such word in a text is

simply treated as a partitioned substring of the input text string. Thus the proper output of the text parser is considered here to be a sequence of sequences of "pretokens", where a pretoken is a text fragment which is a candidate for a word, i.e. an ASCII (sub)string presented for recognition as a system "token". The system lexicon is a lexicographically ordered list of such tokens (with associated type and reference data), and recognition by the system of a pretoken as an actual token is simply a matter of exact string comparison.

Accordingly, the output of the text parser (3.1.2) is a sequence of sequences of pretokens, or sequence of "preexpressions", which is then passed to the type assignment submodule (3.2.1.1) of the type association submodule (3.2.1), where syntactic processing is initiated. Each pretoken is checked against the system lexicon (3.2.0) for its status as a recognized lexical token. If a pretoken is recognized, i.e. if the string comprising that pretoken is included in the lexicon as an actual token (with associated syntactic and semantic data), then it is assigned a lexically associated syntactic type. The system determines at decision node (3.2.1.2) whether all the pretokens from the entered text have been recognized as system tokens. If the determination is negative, then as indicated by the "no" connection to the lexical insertion submodule (3.2.1.3), the user is given the option to add the unrecognized pretokens to the system as tokens with associated type and reference data, i.e. to insert new terms into the lexicon, for further processing. On the other hand, if the determination is affirmative, then the resulting sequence of sequences of lexically typed tokens, or sequence of "virtual" expressions, is passed along the "yes" connection to the type contextualization submodule (3.2.1.4). This submodule initiates a second order type assignment which uses the initial (or virtual) lexical type assignments as data for a contextual process which may reassign these initial types depending on the relative syntactic roles of the tokens in the virtual expressions being processed. Upon complete (re)assignment of appropriate types to tokens, each virtual expression is promoted to an "actual" expression, and each token/type pair becomes a fully functional lexical term with associated semantic data.

Thus the output of the type association submodule (3.2.1) of the syntactic processing module (3.2) comprises a sequence of (actual) expressions, and is passed to the term correlation submodule (3.2.2.1) of the term resolution module (3.2.2). The output of this submodule is a sequence of sequences of fully correlated lexical terms, which is then entered into the term reduction submodule (3.2.2.2), wherein proper syntactic dependencies between terms in an expression are established by means of a type reduction matrix. The output of this submodule is a sequence of sequences of reduction links, which is entered into the term inversion submodule (3.2.2.3), wherein these reduction links are used to construct syntactic trees, each tree representing a processed expression. The resulting sequence of syntactic trees is passed to the syntactic representation submodule (3.2.3), wherein each expression is then represented as a syntactic complex, i.e. a (usually composite) term in the syntactic algebra.

Semantic processing (3.3) is initiated in the semantic representation submodule (3.3.1), wherein the input sequence of syntactic complexes from the syntactic processing module (3.2) is represented as a full semantic complex, i.e. a structure of internal objects in the semantic algebra. This semantic complex is then passed to the formal representation submodule (3.3.2), wherein the input semantic complex is represented as a formal structure adhering to a fundamental transaction paradigm. This formal semantic model is then combined with the sequence of syntactic complexes output from the syntactic processing module to form the input to the formal interpretation submodule (3.3.3), wherein a sequence of formal expressions is constructed as an interpretation of the presented syntactic and semantic data.

In addition, the output of the formal representation submodule is passed to the external representation submodule (3.4.1) of the external processing module (3.4), wherein a specific external representation appropriate for the formal semantic data presented is identified. This external representation is

combined with the sequence of formal expressions output from the formal interpretation submodule to form the input to the external interpretation submodule (3.4.2), wherein a sequence of executable expressions is constructed accordingly for ultimate processing in the appropriate operational environment (3.5).

#### ***DETAILED DESCRIPTION OF THE INVENTION:***

##### ***INTRODUCTION:***

METASCIPT is a translation from a natural language into an executable formal language. This translation is essentially a transformation from the syntactic structures of natural language into effective algebraic forms suitable for further processing. The formal semantics which finally determines the ensuing interpretations and executions of these formal expressions in external operational environments is object-oriented.

The fundamental algorithm upon which METASCIPT is based employs a *reduction* to formal syntactic structures over terms defined in an extensible *lexicon*. This term reduction incorporates both syntactic type and semantic context to achieve an effective formal representation and interpretation of the meaning conveyed by any natural language expression. Extensibility of the lexicon under specific user direction provides the capacity for the system to expand its knowledge of vocabulary and usage, and consequently, offers an effective mechanism under user control for establishing definite incremental enhancements to the system's linguistic capabilities, hence substantially increasing the system's familiarity with (and competence in) particular operational environments. Put simply, the system learns as it goes. In addition, any desired level of syntactic disambiguation is attainable by increasing the local dimensionality of the underlying reduction matrix, though this feature is part of the underlying algorithm, and therefore independent of user modulation.

It should be noted that METASCIPT is not a *speech recognition system*. Instead, it is a fully capable *natural language interpreter*. Specifically, METASCIPT translates natural language expressions into expressions in a formal language associated with an abstract network protocol. A more detailed account of this process follows.

##### ***NOTATION:***

Standard mathematical notation is used to clarify the presentation of certain technical features of the system. In particular, the following set-theoretical notation appears throughout this discussion:

- a) a *set* is a collection of things, called *elements*. For example,  $N = \{0, 1, 2, 3, \dots\}$  is the set of *natural numbers*.

Note: In general, a set  $A$  is most conveniently determined by some property  $P$  of its elements, indicated by use of so-called "set-builder notation" as  $A = \{x \mid P(x)\}$  = the set of things  $x$  satisfying property  $P$ .

- b) the expression ' $x \in A$ ' indicates that the thing  $x$  is an *element* of the set  $A$

c) the expression ' $A \subseteq B$ ' indicates that the set  $A$  is a *subset* of the set  $B$ , i.e. that every element of  $A$  is an element of  $B$  as well

d) a *map* (or *function*) is a relation between two sets  $A, B$  such that each element in  $A$  is assigned a unique element in  $B$ . The expression ' $f: A \rightarrow B$ ' indicates that  $f$  is a map from the set  $A$  to the set  $B$ , i.e.  $f$  assigns a unique element  $y = f(x) \in B$  to each element  $x \in A$ . The *composition* of maps  $f: A \rightarrow B$  and  $g: B \rightarrow C$  on sets  $A, B, C$  is the map  $h = g \circ f: A \rightarrow C$  defined such that  $h(x) = g(f(x)) \in C$  for any  $x \in A$ .

Note: A *program* is a *function* which maps input onto output in an *effective* manner, i.e. by means of a finite, discrete, deterministic procedure; in fact, any process or procedure is *effective* precisely to the extent that it is executable as a *program* of this sort.

e) for any sets  $A, B$ , the *Cartesian product*  $A \times B$  consists of all pairs  $(x, y)$  such that  $x \in A$  and  $y \in B$ , i.e.  $A \times B = \{(x, y) \mid x \in A, y \in B\}$

f) for any sets  $A, B$  the *union*  $A \cup B$  is the set consisting of all elements  $x$  such that  $x \in A$  or  $x \in B$ , i.e.  $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$ ; for any collection  $C = \{A_j \mid 0 \leq j \leq n\}$  of sets  $A_j$  for some  $n \in \mathbb{N}$ , the *overall union*  $\bigcup C$  is the set consisting of unions over all sets  $A_j$ , i.e.  $\bigcup C = \bigcup \{A_j \mid 0 \leq j \leq n\} = A_0 \cup \dots \cup A_n$

g) for any algebras  $A, B$  and representation  $f: A \rightarrow B$ , the *correlated tensor product*  $A \otimes_f B$  is the distinguished subset of  $A \times B$  which consists of all pairs  $(x, f(x))$  for  $x \in A$ , i.e.  $A \otimes_f B = \{(x, f(x)) \mid x \in A\}$  = the *graph* of  $f$ ; for an implicit representation, the map subscript may be omitted, i.e.  $A \otimes B = A \otimes_f B$  for some implicit  $f: A \rightarrow B$

h) for any set  $A$ ,  $\text{Seq}(A)$  is the set of finite *sequences* from  $A$ , i.e.  $\text{Seq}(A) = \{(a_0, \dots, a_n) \mid a_j \in A, n \in \mathbb{N}\}$

#### DEFINITIONS:

*language*: a structure over the following components:

- a) *alphabet*: a set of basic symbols
- b) *punctuation symbols*: a set of basic symbols disjoint from the *alphabet*
- c) *words*: admissible sequences of basic symbols
- d) *punctuations*: admissible sequences of punctuation symbols
- e) *expressions*: admissible sequences of *words* and/or *punctuations*
- f) *sentences*: complete *expressions*
- g) *syntax*: a specification of which
  - sequences of basic symbols are admissible as *words*
  - sequences of punctuation symbols are admissible as *punctuations*
  - sequences of *words* and/or *punctuations* are admissible as *expressions*
  - *expressions* are admissible as *sentences*
- g) *semantics*: a scheme of interpretation over *words* whereby *expressions* acquire meaning with respect to certain external structures

A number of languages enter into this discussion:

- 1) *natural language*: any of the human languages in current use, e.g. English, each characterized by an informal, and hence notoriously ambiguous, syntax and semantics

- 2) *formal language*: a highly structured language, usually mathematical in origin and use, characterized by a uniquely readable, recursive syntax and an extensional, usually first-order semantics; in short, a language for which the syntax and semantics is effectively unambiguous
- 3) *object language*: a formal language which is interpretable relative to a class of extensional structures, i.e. a formal language with an object-oriented semantics
- 4) *protocol language*: a formal language which mediates transactions between addressable nodes on a network
- 5) *executable language*: a formal, programmable language which encodes instructions directly implementable by a suitably capable machine such as a computer

*system*: the integrated process which manifests METASCIPT, and which may be implemented as software running on any programmable device

*string*: a sequence of ASCII characters

*text*: a *string* presented to the *system* as the fundamental unit of initial input

*parser*: a process which partitions *texts* into sequences of sequences of *substrings*

*preexpression*: a sequence of *substrings* of some *text*, distinguished as a unit of syntactic processing by the *text parser*

*lexicon*: a *system* specific, indexed, lexicographically ordered list of designated *strings*, called *tokens*, each of which is associated with certain syntactic and semantic information; in particular, each *token* is associated with a *lexical type*, which may be *virtual* (syntactically ambiguous) or *actual* (syntactically unambiguous); furthermore, each *token* which is associated with an *actual type* is also associated with a *lexical reference*, which provides basic semantic information

Note: A single *string* may serve as a *token* with multiple entries, associated with a number (including 1) of *virtual types* and a number of *actual types*, reflecting that *token*'s multiple syntactic roles, e.g. as a verb and an object, or an object and an adjective, etc. Although there is considerable variability in such syntactic multiplicities among lexical entries, it is still the case that every *token* is associated with at least one *actual type*.

*token*: a *string* recognized by the *system* in the sense that it is included in the *system lexicon*

*type*: a syntactic category used to organize semantically similar *tokens*; there are three sorts:

- 1) *virtual*: a *lexical type* which is ambiguous
- 2) *actual*: a *lexical type* which is not ambiguous
- 3) *reduced*: a *syntactic type* which has specific semantic functionality upon *term reduction*

*term*: there are six sorts:

- 1) *lexical*: a *token/type* pair in the *lexicon* with associated reference data
- 2) *reduced*: a *token/type* pair for which the *type* is *reduced*
- 3) *syntactic*: an element of the *syntactic algebra* associated with a *language*
- 4) *semantic*: an element of the *semantic object algebra* associated with a *language*
- 5) *tensored*: an element of the *semantic tensor algebra* associated with a *language*
- 6) *formal*: an interpretable element of a *formal language*

*reference*: there are two sorts:

- 1) *internal*: an *object* with which a *term* is associated, either in the *lexicon* or in the *semantic object algebra*
- 2) *external*: an *object* with which an internal *semantic object* is associated in some *operational environment*

*expression*: a sequence of *tokens*

*sentence*: a syntactically correct, semantically complete *expression*

*chain*: a linearly ordered set of *nodes* (usually comprising a subset of some *tree*)

*tree*: a partially ordered set of *nodes*

*model*: a semantic structure M for a *formal language* FL consisting of

- a) a *set*  $\text{Dom}(M)$  of individuals (called the “domain” of M)
- b) a *set*  $\text{Rln}(M)$  of relations on  $\text{Dom}(M)$
- c) a *set*  $\text{Obj}(M)$  of objects (each object containing elements of  $\text{Dom}(M)$  as elements; also, there is a *null object*  $0 \in \text{Obj}(M)$  for technical reasons)
- d) a *set*  $\text{Map}(M)$  of functions between objects

These individuals, objects, relations, and functions are formal interpretations of corresponding terms in the language FL, and expressions of FL which correctly describe configurations of these various elements which actually obtain in the model are considered to be “true” in the model, or “satisfied” by the model; this satisfaction relation between a model M and an expression  $\varphi$  of the language FL is denoted as “ $M \models \varphi$ ”, meaning that “M is a model of  $\varphi$ ”, “M satisfies  $\varphi$ ”, “ $\varphi$  is satisfied in M”, “ $\varphi$  is true in M”, or “ $\varphi$  holds in M”.

*operational environment*: a dynamic structure E represented as a series of *static states*  $E_k$  (for  $k \in \mathbb{N}$ ), each of which comprises a *model* for an *executable language* EL

Throughout this discussion, the terms “internal” and “external” are applied relative to the system itself. Thus any component, module, process, or method which is part of the system is considered to be internal, while any user of the system or operational environment for the system is regarded as being external. This distinction is more logical than physical, since an external operational environment for the system may reside on the same computer system or device which hosts the system.

#### **SYSTEM SETS:**

The following *sets* are defined relative to a *language* L:

$\text{Sym}(L)$ : the set of basic *symbols* for L, usually including an *alphabet* and various *punctuation symbols*

$\text{Tok}(L)$ : the set of *lexical tokens* for L, i.e. a distinguished subset of  $\text{Seq}(\text{Sym}(L))$

$Ltp(L)$ : the set of *lexical types* for  $L$

$Rtp(L)$ : the set of *reduced types* for  $L$ , including a *null type*

$Trm(L)$ : the set of *lexical terms* for  $L$ , i.e. a distinguished subset of  $Tok(L) \times Ltp(L)$

$Rdn(L)$ : the set of *reduced terms* for  $L$ , i.e. a distinguished subset of  $Tok(L) \times Rtp(L)$

$Ref(L)$ : the domain of *lexical references* for  $L$

$Lex(L)$ : the *lexicon* for  $L$ , i.e. a distinguished subset of  $N \times Trm(L) \times Ref(L)$  consisting of a lexicographically ordered list of (*index/token/type/reference*) entries

$Txt(L)$ : the set of *texts* for  $L$ , i.e. a subset of  $Seq(Sym(L))$  determined by *user input*

$Prx(L)$ : the set of *preexpressions* for  $L$ , i.e. a subset of  $Seq(Seq(Sym(L)))$  determined by the *parser*

$Exp(L)$ : the set of *expressions* for  $L$ , i.e. a distinguished subset of  $Seq(Tok(L))$

$Snt(L)$ : the set of *sntences* for  $L$ , i.e. a distinguished subset of  $Exp(L)$

$Tre(L)$ : the set of *syntactic trees* for  $L$ , each having *reduced terms* from  $Rdn(L)$  as *nodes*

$Syn(L)$ : the *syntactic algebra* for  $L$

$Sem(L)$ : the *semantic object algebra* for  $L$

$Tns(L)$ : the *semantic tensor algebra* for  $L$ , viz.  $Tns(L) = Syn(L) \otimes Sem(L)$  for a canonical map

$f: Syn(L) \rightarrow Sem(L)$

$Mod(L)$ : the set of *internal formal models* for  $L$

$Env(L)$ : the set of *external operational environments* for  $L$

#### SYSTEM MAPS:

The following maps are defined relative to a *natural language*  $NL$ , an associated *object language*  $XL$ , a *protocol language*  $PL$ , and an *executable language*  $EL$  associated with an operational environment  $E$ :

$txtprs : Txt(NL) \rightarrow Seq(Prx(NL))$  is the *text parser* which maps any *text string*  $s \in Txt(NL)$  onto a sequence  $txtprs(s) \in Seq(Prx(NL))$  of *preexpressions*, each of which is a sequence of *pretokens*, i.e. *strings* which are candidates for *tokens* (including *punctuations*) as determined by inclusion in the *lexicon*

$lextyp : N \times Tok(NL) \rightarrow Typ(NL)$  is the *lexical type assignment* such that  $lextyp(n,a) \in Typ(NL)$  is the *lexical type* associated in the *lexicon* with the *indexed token*  $(n,a)$

$lexref : Tok(NL) \times Typ(NL) \rightarrow Ref(NL)$  is the *lexical reference assignment* such that  $lexref(a,e) \in Ref(NL)$  is the *reference* associated in the *lexicon* with the *token/type* pair  $(a,e) \in Tok(NL) \times Typ(NL)$

Note: If  $lexref(a,e) = \emptyset$  (= *null reference*) then either  $(a,e)$  is not a recognized *token/type* pair, i.e. there is no entry in the *lexicon* of the form  $(n/a/e/r)$  for any index  $n$ , or  $e$  is a *virtual type*; accordingly, the set of *lexical terms* for  $NL$  is defined as  $Trm(NL) = \{(a,e) \in Tok(NL) \times Typ(NL) \mid$

$\text{lexref}(a,e) \neq \emptyset$  }. Thus *lexical reference* properly reduces to an assignment  $\text{lexref} : \text{Trm}(\text{NL}) \rightarrow \text{Ref}(\text{NL})$  on actual *lexical terms*, rather than on random *token/type* pairs, or even on lexical entries in general.

$\text{syntrm} : \text{Rdn}(\text{NL}) \rightarrow \text{Syn}(\text{NL})$  is the assignment of a *basic syntactic term*  
 $q(d,a) = \text{syntrm}(a,d) \in \text{Syn}(\text{NL})$  to each *reduced term*  $(a,d) \in \text{Rdn}(\text{NL})$

$\cdot [\cdot] : \text{Syn}(\text{NL}) \times \text{Syn}(\text{NL}) \rightarrow \text{Syn}(\text{NL})$  is *term application* on pairs of *complexes*  $q, q' \in \text{Syn}(\text{NL})$  such that  $q[q'] \in \text{Syn}(\text{NL})$  is the *composite complex* induced by the application of the syntactic term  $q$  to the term  $q'$

Note: In fact,  $\text{Syn}(\text{NL})$  is simply the algebraic closure under the operation of *term application* over the set of *basic complexes* associated with the *lexical terms* of  $\text{NL}$  by means of the map  $\text{syntrm}$  on  $\text{Rdn}(\text{NL})$ , i.e.  $\text{Syn}(\text{NL})$  is defined by induction on *term application* as follows:

- a) *null complex*: there is a *null term*  $0 \in \text{Syn}(\text{NL})$
- b) *basic complex*:  $q(d,a) = \text{syntrm}(a,d) \in \text{Syn}(\text{NL})$  for any *reduced term*  $(a,d) \in \text{Rdn}(\text{NL})$
- c) *composite complex*:  $q[q'] \in \text{Syn}(\text{NL})$  for any  $q, q' \in \text{Syn}(\text{NL})$ ;  $q[0] = 0[q] = q$  for any  $q \in \text{Syn}(\text{NL})$
- d) *completeness*:  $q \in \text{Syn}(\text{NL})$  iff  $q$  is either the *null complex* or a *basic complex*, or a *composite complex* generated by a finite sequence of *term applications* over a set of *basic complexes*

$\text{syntyp} : \text{Syn}(\text{L}) \rightarrow \text{Rtp}(\text{L})$  is the *syntactic type designator* defined by induction on  $\text{Syn}(\text{L})$  as follows:

- a) *null complex*:  $\text{syntyp}(0) = \text{nul} \in \text{Rtp}(\text{NL})$ , where  $0 \in \text{Syn}(\text{NL})$  is the *null term* and  $\text{nul}$  is the *null type*
- b) *basic complex*:  $\text{syntyp}(q(d,a)) = d \in \text{Rtp}(\text{NL})$  for any *reduced term*  $(a,d) \in \text{Rdn}(\text{NL})$
- c) *composite complex*:  $\text{syntyp}(q[q']) = \text{syntyp}(q) \in \text{Rtp}(\text{NL})$  for any  $q, q' \in \text{Syn}(\text{NL})$ ;

$\text{subtrm} : \text{Syn}(\text{NL}) \times \text{Rtp}(\text{NL}) \rightarrow \text{Syn}(\text{NL})$  is the *subterm designator* such that  $\text{subtrm}(q,d) \in \text{Syn}(\text{NL})$  is the *leading subcomplex* of *syntactic type*  $d \in \text{Rtp}(\text{NL})$  in the *syntactic complex*  $q \in \text{Syn}(\text{NL})$ . The precise definition of  $\text{subtrm}$  proceeds by induction on  $\text{Syn}(\text{NL})$  as follows:

- a) *null complex*:  $\text{subtrm}(0,d) = 0$  for any type  $d \in \text{Rtp}(\text{NL})$ , where  $0 \in \text{Syn}(\text{NL})$  is the *null term*
- b) *basic complex*:  $\text{subtrm}(q(d',a),d) = q(d',a)$  if  $d' = d$ , and  $\text{subtrm}(q(d',a),d) = 0$  if  $d' \neq d$ , for any type  $d \in \text{Rtp}(\text{NL})$  and *reduced term*  $(a,d') \in \text{Rdn}(\text{NL})$
- c) *composite complex*:  $\text{subtrm}(q[q'],d) = \text{subtrm}(q'',d) \in \text{Syn}(\text{NL})$  for any  $q, q' \in \text{Syn}(\text{NL})$  and type  $d \in \text{Rtp}(\text{NL})$ , where  $q'' = q$  if  $\text{subtrm}(q,d) \neq 0$ , and  $q'' = q'$  otherwise

Note:  $\text{subtrm}(q,d) = 0$  if there is no subcomplex of type  $d \in \text{Rtp}(\text{NL})$  in the syntactic complex  $q \in \text{Syn}(\text{NL})$ ; otherwise,  $\text{syntyp}(\text{subtrm}(q,d)) = d \in \text{Rtp}(\text{NL})$ . Also, by default,  $\text{subtrm}(q,\text{nul}) = 0$  for all  $q \in \text{Syn}(\text{NL})$ , i.e. the null term is a subterm of every syntactic complex.

$\text{synrep} : \text{Exp}(\text{NL}) \rightarrow \text{Syn}(\text{NL})$  is the *syntactic representation* such that  $\text{synrep}(p) \in \text{Syn}(\text{NL})$  is the *syntactic complex* associated with the *expression*  $p \in \text{Exp}(\text{NL})$

Note: The *system* currently constructs  $\text{synrep}$  as a composition of the maps  $\text{syntre} : \text{Exp}(\text{NL}) \rightarrow \text{Tre}(\text{NL})$  and  $\text{trerep} : \text{Tre}(\text{NL}) \rightarrow \text{Syn}(\text{NL})$ , where  $\text{syntre}$  associates a *syntactic tree*  $\text{syntre}(p) \in \text{Tre}(\text{NL})$  with each *expression*  $p \in \text{Exp}(\text{NL})$ , and  $\text{trerep}$  associates a *syntactic complex*

$trerep(t) \in \text{Syn}(\text{NL})$  with each *syntactic tree*  $t \in \text{Tre}(\text{NL})$  over the assignment of *basic syntactic terms*  $q(d,a) = \text{syntrm}(a,d)$  to *nodes*  $(a,d) \in t$ , i.e. the representation  $\text{synrep} = \text{trerep} \circ \text{sytre} : \text{Exp}(\text{NL}) \rightarrow \text{Syn}(\text{NL})$

$*$  :  $\text{Sem}(\text{NL}) \times \text{Sem}(\text{NL}) \rightarrow \text{Sem}(\text{NL})$  is the *semantic product* on semantic objects in the *semantic object algebra*  $\text{Sem}(\text{NL})$  such that  $*(x,y) = x*y \in \text{Sem}(\text{NL})$  is the minimal common upper bound of any pair of objects  $x,y$  in the induced class inheritance lattice on  $\text{Sem}(\text{NL})$ , with the second term of the product dominant over the first relative to consistency issues, i.e.  $y$  dominant over  $x$  in  $x*y$

Note: In fact,  $\text{Sem}(\text{NL})$  is simply the algebraic closure under the *semantic product* over the kernel  $\text{Ref}(\text{NL})$  of *lexical references* associated with the *lexical terms* of  $\text{NL}$  by means of the map  $\text{lexref} : \text{Trm}(\text{NL}) \rightarrow \text{Ref}(\text{NL})$ , i.e.  $\text{Sem}(\text{NL})$  is defined by induction on the *semantic product* as follows:

- a) *identity object*: there is an *identity element*  $\mathbf{1} \in \text{Sem}(\text{NL})$
- b) *lexical object*:  $\text{lexref}(a,e) \in \text{Sem}(\text{NL})$  for any *lexical term*  $(a,e) \in \text{Trm}(\text{NL})$ , i.e.  $\text{Ref}(\text{NL}) \subseteq \text{Sem}(\text{NL})$
- c) *composite object*:  $x*y \in \text{Sem}(\text{NL})$  for any objects  $x,y \in \text{Sem}(\text{NL})$ ;  $x*\mathbf{1} = \mathbf{1}*x = x$  for each  $x \in \text{Sem}(\text{NL})$
- d) *completeness*:  $x \in \text{Sem}(\text{NL})$  iff  $x$  is either the *identity object* or a *lexical object*, or a *composite object* generated by a finite sequence of *semantic products* over a set of *lexical objects*

$\text{semrep} : \text{Syn}(\text{NL}) \rightarrow \text{Sem}(\text{NL})$  is the *semantic representation* such that  $\text{semrep}(q) \in \text{Sem}(\text{NL})$  is the *semantic reference* associated with the *syntactic term*  $q \in \text{Syn}(\text{NL})$ ; moreover  $\text{semrep}$  is the implicit representation in the definition of the *semantic tensor algebra*  $\text{Tns}(\text{NL}) = \text{Syn}(\text{NL}) \otimes \text{Sem}(\text{NL}) = \text{Syn}(\text{NL}) \otimes_{\text{semrep}} \text{Sem}(\text{NL})$

Note: *Semantic reference* makes contact with *lexical reference* in that  $\text{semrep}(q(d(e),a)) = \text{lexref}(a,e) \in \text{Ref}(\text{NL})$  for every *basic complex*  $q(d(e),a) \in \text{Syn}(\text{NL})$  naturally associated (after *term reduction*:  $(a,e) \rightarrow (a,d(e))$ ) with a *reduced term*  $(a,d(e)) \in \text{Rdn}(\text{NL})$ ; moreover,  $\text{semrep} : \text{Syn}(\text{NL}) \rightarrow \text{Sem}(\text{NL})$  is a homomorphism from the syntactic algebra to the semantic object algebra in that

- a) *null*  $\rightarrow$  *identity*:  $\text{semrep}(0) = \mathbf{1}$  where  $0 \in \text{Syn}(\text{NL})$  is the *null complex* and  $\mathbf{1} \in \text{Sem}(\text{NL})$  is the *identity object*
- b) *basic*  $\rightarrow$  *lexical*:  $\text{semref}(q(d(e),a)) = \text{lexref}(a,e) \in \text{Sem}(\text{NL})$  for any *lexical term*  $(a,e) \in \text{Trm}(\text{NL})$  and associated *reduced term*  $(a,d(e)) \in \text{Rdn}(\text{NL})$ , where  $q(d(e),a) = \text{syntrm}(a,d(e)) \in \text{Syn}(\text{NL})$  is a *basic complex*
- c) *composite*  $\rightarrow$  *composite*:  $\text{semref}(q[q']) = \text{semref}(q)*\text{semref}(q') \in \text{Sem}(\text{NL})$  for any complexes  $q,q' \in \text{Syn}(\text{NL})$

$\text{tnsrep} : \text{Syn}(\text{NL}) \rightarrow \text{Tns}(\text{NL})$  is the *tensor representation* such that  $\text{tnsrep}(q) = q \otimes \text{semrep}(q) \in \text{Tns}(\text{NL})$  is the *tensored complex* associated with the *syntactic term*  $q \in \text{Syn}(\text{NL})$ , and implicitly, with the *semantic reference*  $\text{semrep}(q) \in \text{Sem}(\text{NL})$

$\text{fmlrep} : \text{Sem}(\text{NL}) \rightarrow \text{Mod}(\text{XL})$  is the *formal representation* such that  $\text{fmlrep}(x) \in \text{Mod}(\text{XL})$  is the *formal model* associated with the *semantic context*  $x \in \text{Sem}(\text{NL})$

$\text{fmlint} : \text{Tns}(\text{NL}) \rightarrow \text{Exp}(\text{XL})$  is the *formal interpretation* such that  $\text{fmlint}(u) \in \text{Exp}(\text{XL})$  is the *formal*

*expression associated with the tensored complex  $u \in Tns(NL)$*

*modref<sub>M</sub> : Trm(XL) → M* for each  $M \in Mod(XL)$  is the *model reference* such that  $modref_M(v) \in M$  is the *model element* referenced by the term  $v \in Trm(XL)$ ; if there is no actual reference for  $v$  in  $M$  then  $modref_M(v) = 0 \in M$  is the *null object* by default

*extrep : Mod(XL) → Env(NL)* is the *external representation* such that  $extrep(M) \in Env(NL)$  is the *external operational environment* associated with the *internal formal model*  $M \in Mod(XL)$

*extint<sub>E</sub> : Trm(PL) → Trm(EL)* is the *external term interpretation* relative to an *operational environment*  $E \in Env(NL)$  with an associated *executable language*  $EL$  such that  $extint_E(v) \in Trm(EL)$  is the *external term* corresponding to the *protocol term*  $v \in Trm(PL)$

*exttrn<sub>E</sub> : Exp(XL) → Exp(EL)* is the *external translation* relative to an *operational environment*  $E \in Env(NL)$  with an associated *executable language*  $EL$  such that  $exttrn_E(\varphi) \in Exp(EL)$  is the *executable translate* of the *formal expression*  $\varphi \in Exp(XL)$

*envexc<sub>E</sub> : Exp(EL) → E* is the *execution process* for an *operational environment*  $E \in Env(NL)$  with an associated *executable language*  $EL$  such that  $envexc_E(\xi) \in E$  is the result of executing the *formal expression*  $\xi \in Exp(EL)$  in  $E$

Note: An *execution process* for an *operational environment*  $E = \{E_k \mid k \in N\}$  and *executable language*  $EL$  is defined with respect to an *execution procedure*  $envprc_E : Exp(EL) \times N \rightarrow N$  such that for any *executable expression*  $\xi \in Exp(EL)$  and *operational state index*  $k \in N$ , the *image state*  $E_{J(k)} \models \xi$  (i.e. the model  $E_{s(k)}$  satisfies the expression  $\xi$ ), where the *image index*  $J(k) = envprc_E(\xi, k) > k$ . In terms of a sequence of operations in  $E$ , any specific *execution* of an expression  $\xi \in Exp(EL)$  is then simply the *operational state*  $envexc_E(\xi) = envprc_E(\xi, n) \in E$  for some index  $n \in N$ .

*pclint : Sem(NL) → Trm(PL)* is the *protocol interpretation* such that  $pclint(x) \in Trm(PL)$  is the *protocol term* corresponding to the *semantic object*  $x \in Sem(NL)$

*pclcod : Mod(XL) × Exp(XL) → Exp(PL)* is the *protocol encoding* such that  $pclcod(M, \varphi) \in Exp(PL)$  is the *protocol expression* associated with the *formal expression*  $\varphi \in Exp(XL)$  as interpreted with respect to the *formal model*  $M \in Mod(XL)$

*pclrep : Exp(PL) → Env(NL)* is the *external protocol representation* such that  $pclrep(X) \in Env(NL)$  is the *external operational environment* encoded into the *protocol expression*  $X \in Exp(PL)$

*pcltrn : Exp(PL) → Exp(EL)* is the *external protocol translation* such that  $pcltrn(X) \in Exp(EL)$  is the *formal expression* encoded into the *protocol expression*  $X \in Exp(PL)$  for execution in the *external operational environment*  $E = pclrep(X) \in Env(NL)$ .

*extref : Exp(PL) × Trm(PL) → Env\*(NL) = Env(NL) ∪ (UEnv(NL))* is the uniform *external reference* such that  $extref(X, T)$  is the external element referenced by the term of *metatype*  $T \in Typ(PL)$  in the *protocol expression*  $X \in Exp(PL)$ ; if there is no actual reference of this type then  $extref(X, T) = 0$  is

the *null object* by default

Note: The uniform range  $\text{Env}^*(\text{NL}) = \text{Env}(\text{NL}) \cup (\text{Env}(\text{NL}))$  of  $\text{extref}$ , where  $\text{UEnv}(\text{NL}) =$  the combined set of elements from all environments  $E \in \text{Env}(\text{NL})$ , accommodates reference with respect to terms of *metatype*  $\text{ENV} \in \text{Typ}(\text{PL})$  as well as all other terms, since the reference  $\text{extref}(X, \text{ENV})$  is an *operational environment* for any protocol expression  $X \in \text{Exp}(\text{PL})$ , i.e.  $\text{extref}(X, \text{ENV}) \in \text{Env}(\text{NL})$ , whereas the referenced external element  $\text{extref}(X, \text{ABC})$  for any *other* metatype  $\text{ABC} \in \text{Typ}(\text{PL})$  is an element of the operational environment  $\text{extref}(X, \text{ENV})$ , i.e.  $\text{ABC} \neq \text{ENV} \Rightarrow \text{extref}(X, \text{ABC}) \in \text{extref}(X, \text{ENV})$ ; therefore, the range of  $\text{extref}$  must be mixed between operational environments and elements of these environments.

$\text{envref}_E : \text{Trm}(\text{PL}) \rightarrow E$  for each  $E \in \text{Env}(\text{NL})$  is the *operational environment reference* such that  $\text{envref}_E(v) \in E$  is the *environment element* referenced by the term  $v \in \text{Trm}(\text{PL})$ ; if there is no actual reference for  $v$  in  $E$  then  $\text{envref}_E(v) = 0 \in E$  is the *null object* by default

Put succinctly, METASCIPT is an effective transformation  $\text{mscript} : \text{Exp}(\text{NL}) \rightarrow \text{Exp}(\text{PL})$  of natural language expressions into formal expressions in the abstract language  $\text{PL}$  associated with a universal protocol. Using a formal metasemantics over the object language  $\text{XL}$  associated with  $\text{NL}$ , this protocol is explicitly designed to accommodate further effective translations  $\text{exttrn}_E : \text{Exp}(\text{XL}) \rightarrow \text{Exp}(\text{EL})$  for specific operational environments  $E \in \text{Env}(\text{NL})$  with associated executable languages  $\text{EL}$ .

It should be noted that a particular protocol is utilized in certain sections of this description for the purposes of specificity and clarity. In this exemplary embodiment, the protocol is XMP (for eXternal Media Protocol), which is designed and used as a universal transaction medium for diverse digital components in a networked environment. However, the invention and functionality of this system is not limited to any specific protocol.

#### **SYSTEM PROCESSES:**

*Text Processing:* Refer to module 3.1 of FIG. 3.

*Text Input:* Refer to submodule 3.1.1 of FIG. 3.

Any *text* is presented to the system as a string of ASCII characters. The specific method of text presentation is irrelevant. It may be voice-generated, entered via keyboard, optically scanned, etc. The *system* itself does not explicitly include a module for speech recognition, optical character recognition, etc., but any effective mechanism for presenting ASCII text to the system is sufficient for its successful operation.

The following text string represents a typical input:

Send Bob an email asking him if he is going to go to his appointment by himself.

This example will be carried throughout the following discussion in order to illustrate the METASCIPT process in some detail.

*Text Parser:* Refer to submodule 3.1.2 of FIG. 3.

The text parser partitions any string of ASCII characters into sequences of sequences of ASCII substrings, using embedded blank spaces as delimiters. These inner sequences of substrings represent potential *expressions*, and the substrings themselves represent potential words or punctuation marks. For terminological convenience, any such parsed substring is called a *pretoken*, and the partitioned sequence in which it occurs is called a *preexpression*.

Applied to the sample text string introduced above, the parser produces the following sequence of pretokens:

```
(send, Bob, an, email, asking, him, if, he, is, going, to, go, to, his,  
appointment, by, himself,..)
```

*Syntactic Processing:* Refer to module 3.2 of FIG. 3.

*Type Association:* Refer to submodule 3.2.1 of FIG. 3.

*Type Assignment:* Refer to submodule 3.2.1.1 of FIG. 3.

Each parsed *pretoken* is checked against the *lexicon* for recognition by the *system*. If a *pretoken* is recognized, i.e. if that string is included in the *lexicon* as an actual *token*, then it is immediately associated with some *type*. This first order assignment of type to token is only tentative at this point in the process, since correct type association requires more than mere lexical recognition of text strings as legitimate *tokens*. Accordingly, these initially assigned types are considered to be “virtual” types.

Virtual type assignment on the sample preexpression parsed above yields the following list of token/type pairs (lexical terms) in the form  $(a, \text{lextyp}(0, a))$  for each indicated token  $a \in \text{Tok}(\text{NL})$  as shown in FIG. 4a. Since all pretokens listed there are recognized by the system as actual tokens, the parsed preexpression becomes an expression for further processing. However, the presence of ambiguous (virtual) types classifies this expression as “virtual”. Promotion to an “actual” expression is deferred to the *type contextualization* process.

*Lexical Insertion:* Refer to submodule 3.2.1.3 of FIG. 3.

If a pretoken is not recognized by the system, then the user is prompted for information concerning lexical type and reference which may be properly associated with that pretoken in order to form a lexical term appropriate for inclusion in the lexicon. Upon such inclusion, the pretoken becomes a system token. This is the primary mechanism, and the only one under direct user control, by which the system learns new vocabulary.

*Type Contextualization:* Refer to submodule 3.2.1.4 of FIG. 3.

Second order type assignment uses the initial lexical type assignments to tokens in an expression as data for a contextual process by which actual types may be assigned to tokens depending on their syntactic roles relative to other tokens in that expression. For example, in the sentence

I want to go to the store.

the word “to” is of ambiguous type because it appears in two different guises, once as the prefix of the infinitive “to go” and later as the preposition in the phrase “to the store”. The system is able to discern such grammatical differences and assign correct types based on syntactic context.

This method of type reassignment through syntactic context alone represents the simplest, most direct form of disambiguation employed by the system. More subtle mechanisms for further disambiguation, such as local type reduction and semantic contextualization, are deployed later in the process.

In any case, final type association on the sample expression being processed yields the list shown in FIG. 4b of refined lexical terms in the form  $(a, \text{lextyp}(j, a))$  for each indicated token  $a \in \text{Tok}(\text{NL})$ , where  $j \geq 0$  is the (perhaps alternative) index corresponding to the appropriate refined type. Note that all the ambiguous (virtual) lexical types have been replaced by unambiguous (actual) types. This type reassignment promotes a *virtual expression* to an *actual expression* suitable for further syntactic processing.

*Term Resolution:* Refer to submodule 3.2.2 of FIG. 3.

*Term Correlation:* Refer to submodule 3.2.2.1 of FIG. 3.

Indirect references by various syntactic elements such as pronouns must be correlated to direct references elsewhere in the text. This task is achieved through type matching in the context of appropriate syntactic configurations. For example, in the sentence

Inform Bob about the next meeting, and tell him that it will happen later than usual.

the pronoun “him” naturally correlates with “Bob”, and “it” with “meeting”. The system establishes such correlations first by executing a simple type/reference matching such as  $(\text{him} \rightarrow \text{male person} \rightarrow \text{Bob})$  on antecedents, and then by evaluating such matches for probable fit according to context. For example, in the extended sentence

Inform Bob about the next meeting *at the factory*, and tell him that it will happen later than usual.

there are two possible type/reference matches for “it”, viz.  $(\text{it} \rightarrow \text{some object} \rightarrow \{\text{meeting}, \text{factory}\})$ , but clearly, on the basis of object characteristics, the match  $(\text{it} \rightarrow \text{some object} \rightarrow \text{meeting})$  is a better fit than  $(\text{it} \rightarrow \text{some object} \rightarrow \text{factory})$  in the context of the phrase “...it will happen later than usual” since meetings happen more readily than factories do, as any capable scheme of lexical reference will indicate. It should be noted, of course, that this process must be applied to a text as a whole, not just

to individual expressions, so that indirect references across multiple expressions may be correlated properly.

Accordingly, the following simple term correlations are made within the sample expression being processed:

- 5) (him, ppm) → (Bob, pnm)
- 7) (he, ppm) → (Bob, pnm)
- 13) (his, psm) → (Bob's, psm)
- 16) (himself, prm) → (Bob, pnm)

*Term Reduction:* Refer to submodule 3.2.2.2 of FIG. 3.

Proper syntactic dependencies between terms in an expression are established by means of a *type reduction* matrix. The dimensions of this matrix help to determine the level of syntactic disambiguation quickly achievable by the system. A 2-dimensional matrix, which maps pairs of tokens into their relative reduction ordering, is minimal. Higher dimensional reduction matrices, which map longer sequences of tokens into their relative reduction orderings, are increasingly effective, but more costly to implement in terms of memory requirements and processing speed. An optimal number of reduction dimensions, of course, depends critically on a complex combination of implementation constraints and performance criteria. Whatever the number of dimensions used, however, the system is designed to establish correct syntactic relationships on a relatively global scale (at least over an entire expression) simply by executing local term reductions in the proper order.

For example, using a minimal matrix on the sample expression being processed, the local term reduction sequence is constructed as shown in FIG. 5.

*Term Inversion:* Refer to submodule 3.2.2.3 of FIG. 3.

Proper chains of syntactic dependencies among tokens in a sentence, and the resultant dependencies between those chains, are constructed by means of an effective inversion of the term reduction process. These chains are then used to generate branches of the syntactic tree which represents the underlying syntactic structure of the expression being processed.

On the sample reduction sequence just constructed, term inversion produces the maximal chains shown in FIG. 6a. Note that term reduction has effected critical type modulations in some of the subordinate terms, viz. obj→obd (general object → direct object), obj→obi (general object → indirect object), obj→obs (general object → verb subject), and obj→obp (general object → preposition object). These reduced types are critical data for accurate semantic processing.

*Syntactic Representation:* Refer to submodule 3.2.3 of FIG. 3.

Each expression for a language is represented at the fundamental syntactic structural level by a tree, i.e. by a finitely branching partial order of finite length having elements corresponding to lexical

terms ordered by their associated type reductions. This tree has a canonical representation as a natural branching diagram, which in turn becomes represented as a (usually composite) term (*syntactic complex*) in an associated syntactic algebra. The branches of this tree correspond directly to the chains of syntactic dependencies established in the *term resolution* process.

For example, the syntactic tree representing the sample text being processed becomes structured as shown in FIG. 6b. Nodes of a syntactic tree and the syntactic order relations which hold between them are interpreted in terms of a many-sorted syntactic structure. The canonical language for this class of structures is based on the notion of a type/token *complex* to form basic terms, and the operation of *term application* on pairs of these complexes to form more complicated or composite terms. Each term, whether basic or composite, ultimately corresponds to a semantic object with certain characteristics, and these objects in their various configurations comprise the domains and internal relations of the syntactic structures considered.

More specifically, under METASCIPT the formal correlate of an expression in a natural language is a term in the associated syntactic algebra which represents the effective translation of the expression into a form suitable for direct system interpretation. This algebra is based on an operation of *term application* which, given a translation of natural language expressions into algebraic terms, transforms syntactic dependencies at the level of natural language into specific formal relations at an algebraic level. Thus the syntax for natural language becomes a matter of effective computation.

Each node (type/token pair)  $(d,a)$  in a syntactic tree  $t(p) = \text{syntre}(p) \in \text{Tre}(\text{NL})$  representing an expression  $p \in \text{Exp}(\text{NL})$  corresponds to a reduced term (token/type pair)  $(a,d) \in \text{Rdn}(\text{NL})$ , and is associated with a basic complex  $q(d,a) \in \text{Syn}(\text{NL})$  by means of the syntactic assignment  $\text{syntrm}$  :  $\text{Rdn}(\text{NL}) \rightarrow \text{Syn}(\text{NL})$ , i.e.  $q(d,a) = \text{syntrm}(a,d)$ . The immediate syntactic dependence of another term  $(d',a')$  on  $(d,a)$  determined by term reduction is signified at the tree level as the order relation  $(d,a) \text{ L } (d',a')$ . The algebraic operation of *term application*  $\cdot[\cdot] : \text{Syn}(\text{NL}) \times \text{Syn}(\text{NL}) \rightarrow \text{Syn}(\text{NL})$  on pairs of syntactic complexes then yields composite complexes of the form  $q'' = q[q']$  induced by these syntactic dependency relations; for example, in the particular case considered here involving the terms  $(d,a)$  and  $(d',a')$ , term application results in the composite complex  $q''((d,a),(d',a')) = q(d,a)[q'(d',a')] = \text{syntrm}(a,d)[\text{syntrm}(a',d')]$ .

Similarly, a chain of syntactic dependencies yields iterated applications on increasingly composite complexes and the basic complex which represents the next link in the dependency chain, i.e. the relation  $z_0 \text{ L } z_1 \text{ L } z_2$  between syntactic terms  $z_0, z_1, z_2 \in \text{Trm}(\text{NL})$  yields the algebraic term  $q_0[q_1[q_2]] \in \text{Syn}(\text{NL})$  where  $q_j = \text{semtrm}(z_j)$ , etc. On the other hand, branching in a syntactic tree yields a slightly different form for the associated algebraic term, in that a dependency relation of the form  $z_0 \text{ L } (z_1, z_2)$ , whereby syntactic terms  $z_1$  and  $z_2$  are both directly dependent on  $z_0$  (i.e.  $z_0 \text{ L } z_1$  and  $z_0 \text{ L } z_2$ ), yields the algebraic term  $q_0[q_1][q_2] = (q_0[q_1])[q_2] \neq q_0[q_1[q_2]]$ , i.e term application is not an *associative* algebraic operation.

In any case, continued term applications of this sort, as explicitly induced by the dependency structure of the syntactic tree, thus yield an effective representation of any expression in a language  $\text{NL}$  as a (usually composite) term in the associated syntactic algebra  $\text{Syn}(\text{NL})$ . For example, the syntactic complex corresponding to the sample expression being processed is

```
(act, send) [ (obi, Bob) ] [ (obd, email) [ (adj, an) ] [ (ptc, asking)
[ (ltm, if) [ (act, is) [ (ptc, going) [ (inf, to) [ (act, go)
```

[(prp,to) [(obp,appointment) [(adj,his)]]] ]  
[(prp,by) [(obp,himself) ]]]]]]]]]]

This algebraic form of the expression is critical for the language processing which follows; in particular, the syntactic representation of expressions afforded by algebraic term reduction provides an effective recursion structure for accurate semantic processing.

In summary, each expression  $p \in \text{Exp}(\text{NL})$  is represented as a syntactic tree  $t(p) = \text{syntre}(p) \in \text{Tre}(\text{NL})$  which induces an associated syntactic complex  $q(t(p)) = \text{trerep}(t(p)) \in \text{Syn}(\text{NL})$ . A detailed definition of the intermediate syntactic algebraic representation  $\text{trerep} : \text{Tre}(\text{NL}) \rightarrow \text{Syn}(\text{NL})$  is given by induction on syntactic dependence:

- a) *null dependence*:  $trerep(d,a) = syntrm(a,d)$  for any trivial (single node) tree  $(d,a) \in Tre(NL)$
- b) *direct dependence*:  $trerep(t \sqcup t') = trerep(t)[trerep(t')]$  for any trees  $t,t' \in Tre(NL)$
- c) *multiple dependence*:  $trerep(t \sqcup (t',t'')) = (trerep(t)[trerep(t')])[trerep(t'')]$  for any trees  $t,t',t'' \in Tre(NL)$

Full syntactic representation is then simply the composition  $synrep = trerepo \circ syntre : \text{Exp}(\text{NL}) \rightarrow \text{Syn}(\text{NL})$ .

*Semantic Processing:* Refer to module 3.3 of FIG. 3.

*Semantic Representation:* Refer to submodule 3.3.1 of FIG. 3.

The lexical reference map  $lexref : \text{Trm}(\text{NL}) \rightarrow \text{Ref}(\text{NL})$  forms the basis for the interpretation of terms in the semantic algebra  $\text{Sem}(\text{NL})$  as objects constructed over the reference domain  $\text{Ref}(\text{NL})$ .

Specifically, each lexical term (token/type pair)  $(a,e) \in \text{Trm}(\text{NL})$  is explicitly associated in the *lexicon* with a reference  $\text{lexref}(a,e) \in \text{Ref}(\text{NL})$  which instantiates the term in a given environment; in fact, no lexical term is properly defined until such a reference for it is specified, since this definition forms the principal link between a natural language term and its intended meanings. The basic objects of the semantic algebra are simply these lexical references, i.e.  $\text{Ref}(\text{NL}) \subseteq \text{Sem}(\text{NL})$ .

This first order notion of reference for lexical terms is then extended to more complex semantic terms by means of a semantic product  $* : \text{Sem}(\text{NL}) \times \text{Sem}(\text{NL}) \rightarrow \text{Sem}(\text{NL})$  on objects which allows a proper definition of a semantic representation  $\text{semrep} : \text{Syn}(\text{NL}) \rightarrow \text{Sem}(\text{NL})$  on the entire algebra  $\text{Syn}(\text{NL})$ . By induction on the composition of syntactic complexes, the full definition becomes:

- a) *null*:  $semrep(0) = 1$  where  $0 \in \text{Syn}(\text{NL})$  is the *null complex* and  $1 \in \text{Sem}(\text{NL})$  is the *identity object*
- b) *basic*:  $semref(q(d(e),a)) = lexref(a,e) \in \text{Sem}(\text{NL})$  for any *lexical term*  $(a,e) \in \text{Trm}(\text{NL})$  and associated *reduced term*  $(a,d(e)) \in \text{Rdn}(\text{NL})$ , where  $q(d(e),a) = syntrm(a,d(e)) \in \text{Syn}(\text{NL})$  is a *basic complex*
- c) *composite*:  $semref(q[q']) = semref(q)*semref(q') \in \text{Sem}(\text{NL})$  for any complexes  $q,q' \in \text{Syn}(\text{NL})$

By this definition,  $\text{semrep}$  is clearly a homomorphism from the syntactic algebra into the semantic object algebra.

For example, given the dependency structure  $z_0 \sqsubset z_1 \sqsubset z_2$  on certain lexical terms  $z_j = (a_j, e_j) \in \text{Trm}(\text{NL})$  with associated syntactic complexes  $q_j = q(z_j) \in \text{Syn}(\text{NL})$  for  $j=0,1,2$ , there is a composite complex  $q_3 = q_0[q_1[q_2]] \in \text{Syn}(\text{NL})$  constructed by iterated syntactic term application with semantic reference

$$\begin{aligned} \text{semrep}(q_3) &= \text{semrep}(q_0[q_1[q_2]]) \\ &= \text{semrep}(q_0) * \text{semrep}([q_1[q_2]]) \\ &= \text{semrep}(q_0) * (\text{semrep}(q_1) * \text{semrep}(q_2)) \\ &= \text{lexref}(a_0, e_0) * (\text{lexref}(a_1, e_1) * \text{lexref}(a_2, e_2)) \end{aligned}$$

in  $\text{Sem}(\text{NL})$  derived from the individual lexical references  $\text{lexref}(a_j, e_j) \in \text{Ref}(\text{NL}) \subseteq \text{Sem}(\text{NL})$ .

The definition of the semantic product  $* : \text{Sem}(\text{NL}) \times \text{Sem}(\text{NL}) \rightarrow \text{Sem}(\text{NL})$  is based on a notion similar to class inheritance from the formal practice of object-oriented design (OOD). Specifically, each reference in  $\text{Sem}(\text{NL})$  is instantiated as an object with certain characteristics, and the product  $x * y$  of two objects  $x, y \in \text{Sem}(\text{NL})$  is simply the object  $z \in \text{Sem}(\text{NL})$  generated in the inheritance lattice as the minimal common upper bound of  $x$  and  $y$ , with  $y$  dominant over  $x$  on issues of conjunctive consistency. Note that by virtue of consistency dominance this product need not be *commutative*, i.e. it is not necessarily the case that  $x * y = y * x$  for all  $x, y \in \text{Sem}(\text{NL})$ ; similarly, this product need not be *associative*, i.e. it is not necessarily the case that  $x * (y * z) = (x * y) * z$  for all  $x, y, z \in \text{Sem}(\text{NL})$ . However, it is *idempotent*, i.e.  $x * x = x$  for all  $x \in \text{Sem}(\text{NL})$ .

It is primarily through this algebraic generation of composite objects that the system gains its complexity; moreover, the addition of terms to the lexicon which have these composite objects as their direct lexical references permits unlimited efficiency and sophistication of machine comprehensible natural language usage.

The semantic tensor algebra  $\text{Tns}(\text{NL}) = \text{Syn}(\text{NL}) \otimes \text{Sem}(\text{NL})$  defined over the representation  $\text{semrep} : \text{Syn}(\text{NL}) \rightarrow \text{Sem}(\text{NL})$  is composed of tensored correlations  $q \otimes \text{semrep}(q)$  of syntactic terms  $q \in \text{Syn}(\text{NL})$  and associated semantic representations  $\text{semrep}(q) \in \text{Sem}(\text{NL})$ , which form the direct computational basis for formal interpretations of natural language expressions in terms of a fundamental transaction paradigm. Specifically, the syntactic relationships encoded in a syntactic complex  $q(p) \in \text{Syn}(\text{NL})$  derived from an expression  $p \in \text{Exp}(\text{NL})$  permit an exact internal structuring of the associated semantic complex  $\text{semrep}(q(p)) \in \text{Sem}(\text{NL})$ .

*Formal Representation:* Refer to submodule 3.3.2 of FIG. 3.

Associated with any expression  $p \in \text{Exp}(\text{NL})$  is the *semantic context*  $c(p) = \text{contxt}(q(p), s(p)) \in \text{Sem}(\text{NL})$  where  $q(p) = \text{synrep}(p) \in \text{Syn}(\text{NL})$  is the syntactic complex representing  $p$  and  $s(p) \in \text{Txt}(\text{NL})$  is the original input text from which  $p$  is derived through text parsing and type association. This semantic context  $c(p)$  is formally represented as an internal structure  $M(p) = \text{fnlrep}(c(p)) \in \text{Mod}(\text{XL})$ , which serves as an abstract model of the operational environment in which  $p$  is to be executed after translation into a suitably effective form. As formal representatives in this modeling capacity, these internal structures for  $\text{XL}$  form the critical link between  $\text{NL}$  and any executable language  $\text{EL}$  for an external environment  $\text{E}$ .

The *semantic context*  $c(p)$  is constructed as an object in the algebra  $\text{Sem}(\text{NL})$  as follows: After text parsing and type association, the text  $s(p)$  is represented as a sequence  $(p_0, \dots, p_n) \in \text{Seq}(\text{Exp}(\text{NL}))$  of expressions  $p_j \in \text{Exp}(\text{NL})$  for some  $n \geq 0$ , where  $p = p_j(p)$  for some index  $j(p) \in \{0, \dots, n\}$ . Note that  $s(p_j) = s(p)$  for  $0 \leq j \leq n$ . The definition of the representation  $\text{contxt}(q(p_j), s(p)) \in \text{Sem}(\text{NL})$  proceeds by induction on the indices  $j \in \{0, \dots, n\}$  where  $q(p_j) \in \text{Syn}(\text{NL})$  is the syntactic complex representing  $p_j$ . Entering into this definition is the *metaterm* operator  $\text{pcltrm} : \text{Syn}(\text{NL}) \times \text{Typ}(\text{PL}) \rightarrow \text{Syn}(\text{NL})$  associated with the protocol language PL as described in the following section. Briefly, for any syntactic complex  $q \in \text{Syn}(\text{NL})$  and *metatype*  $ABC \in \text{Typ}(\text{PL})$ , the syntactic term  $\text{pcltrm}(q, ABC) \in \text{Syn}(\text{NL})$  is the leading subterm of  $q$  of type  $ABC$ ; in particular, subterms  $\text{pcltrm}(q, \text{ENV}) \in \text{Syn}(\text{NL})$  of *metatype*  $\text{ENV} \in \text{Typ}(\text{PL})$  (indicating “environment”) play a significant role.

Specifically, the inductive definition of  $\text{contxt} : \text{Syn}(\text{NL}) \times \text{Txt}(\text{NL}) \rightarrow \text{Sem}(\text{NL})$  is as follows:

- a)  $j = 0: \text{contxt}(q(p_0), s(p)) = \text{semrep}(\text{pcltrm}(q(p_0), \text{ENV})) \in \text{Sem}(\text{NL})$
- b)  $j = k+1$  for  $k < n: \text{contxt}(q(p_{k+1}), s(p)) = \text{contxt}(q(p_k), s(p)) * \text{semrep}(\text{pcltrm}(q(p_{k+1}), \text{ENV})) \in \text{Sem}(\text{NL})$

Then  $c(p) = \text{contxt}(q(p), s(p)) = \text{contxt}(q(p_j(p)), s(p)) \in \text{Sem}(\text{NL})$ , and  $M(p) = \text{fmlrep}(c(p)) \in \text{Mod}(\text{XL})$  by means of the *formal representation*  $\text{fmlrep} : \text{Sem}(\text{NL}) \rightarrow \text{Mod}(\text{XL})$ . The noncommutativity of the semantic product is critical in this definition.

*Formal Interpretation:* Refer to submodule 3.3.3 of FIG. 3.

The formal term interpretation  $\text{trmint} : \text{Sem}(\text{NL}) \times \text{Mod}(\text{XL}) \rightarrow \text{Trm}(\text{XL})$  establishes the syntactic basis for the formalization of NL, since expressions  $\varphi(p) \in \text{Exp}(\text{XL})$  constructed as formal interpretations of natural expressions  $p \in \text{Exp}(\text{NL})$  are built from terms in  $\text{Trm}(\text{XL})$  associated through  $\text{trmint}$  with appropriate objects in  $\text{Sem}(\text{NL})$ . In some extremely approximate sense,  $\varphi(p) = q(p) \otimes \text{trmint}(x(p))$  where  $q(p) = \text{synrep}(p) \in \text{Syn}(\text{NL})$  is the syntactic complex representing  $p$  and  $x(p) = \text{semrep}(q(p)) \in \text{Sem}(\text{NL})$  is the semantic complex representing  $q(p)$ .

The formal semantics, or metasemantics, by which these formal interpretations are constructed is called MEAO (for Modified Environment Action Object). It is based on a fundamental transaction paradigm for which the elementary operation is a mapping  $f : A \rightarrow B$  on structures  $A, B \in \text{Obj}(M)$  in an environment  $M$ . Terms of an object language for this semantics reflect this fundamental orientation, and a basic statement in any such language is simply an assignment of the form “ $x_B = f_B^A(x_A)$ ” where  $x_A \in A$ ,  $x_B \in B$ , and  $f_B^A : A \rightarrow B$ . More complicated expressions are constructed from these basic (or atomic) statements by means of the usual formal propositional connectives (negation, conjunction, disjunction, and implication) and quantification (existential and universal). Of course, any of the individuals, objects, functions, and environments which serve as interpreted elements for such a language may be arbitrarily articulated, thus extremely complex despite the apparent simplicity of this formal syntax; hence the notion of “modification” (the  $M$  in MEAO) plays a significant role.

As previously indicated, the proper formal interpretation of an expression  $p \in \text{Exp}(\text{NL})$  relative to a semantics for the object language  $\text{XL}$  is induced by a map  $\text{trmint} : \text{Sem}(\text{NL}) \rightarrow \text{Trm}(\text{XL})$ . However, the actual computation of MEAO elements from the syntactic and semantic apparatus of  $\text{NL}$  is performed in terms of a formal *protocol* which implements this metasemantics in an effective manner.

Specifically, there is a *metaterm* operator  $\text{pcltrm} : \text{Syn}(\text{NL}) \times \text{Typ}(\text{PL}) \rightarrow \text{Syn}(\text{NL})$  for the protocol language  $\text{PL}$  associated with  $\text{XL}$ , which is built from conditional sequences (in essence, executable routines) of syntactic subterm assignments yielded by the standard *subterm designator subtrm* :  $\text{Syn}(\text{NL}) \times \text{Rtp}(\text{NL}) \rightarrow \text{Syn}(\text{NL})$ . Distinguished semantic types  $T \in \text{Typ}(\text{PL})$  determine certain syntactic subterms  $q(p)_T = \text{pcltrm}(q(p), T) \in \text{Syn}(\text{NL})$  of the syntactic complex  $q(p) \in \text{Syn}(\text{NL})$ , and these syntactic subterms yield semantic representations  $x(p)_T = \text{semrep}(q(p)_T) \in \text{Sem}(\text{NL})$ .

It is at this point that the formal term interpretation  $\text{trmint} : \text{Sem}(\text{NL}) \rightarrow \text{Trm}(\text{XL})$  comes into play as the basis of the formal semantics for  $\text{XL}$ . From the perspective of an effective metasemantics for  $\text{XL}$ , however, this term interpretation is derived from a higher level *protocol interpretation pclint* :  $\text{Sem}(\text{NL}) \rightarrow \text{Trm}(\text{PL})$  by composition with an intermediate formal interpretation  $\text{fmltrm} : \text{Trm}(\text{PL}) \rightarrow \text{Trm}(\text{XL})$  as

$$\text{trmint} = \text{fmltrm} \circ \text{pclint} : \text{Sem}(\text{NL}) \rightarrow \text{Trm}(\text{PL}) \rightarrow \text{Trm}(\text{XL})$$

It is in this precise sense, coupled with the fact that the internal model  $M(p) = \text{fmlrep}(c(p)) \in \text{Mod}(\text{XL})$  is also interpreted as an element for  $\text{PL}$  corresponding to a term  $V(p)_{\text{ENV}} = \text{pclint}(c(p), \text{ENV}) \in \text{Trm}(\text{PL})$  relative to the *metatype*  $\text{ENV} \in \text{Typ}(\text{PL})$  (indicating “environment”), that  $\text{PL}$  constitutes an effective *formal* metalanguage for  $\text{XL}$ , whereas  $\text{NL}$  is an *informal* metalanguage for  $\text{XL}$ .

In any case, the semantic representations  $x(p)_T \in \text{Sem}(\text{NL})$  are interpreted as formal terms  $v(p)_T = \text{trmint}(x(p)_T) = \text{fmltrm} \circ \text{pclint}(x(p)_T) \in \text{Trm}(\text{XL})$ . These terms then designate formal elements  $h(p)_T = \text{modref}_M(p)(v(p)_T) \in M(p)$  by means of the model reference map  $\text{modref}_M(p) : \text{Trm}(\text{XL}) \rightarrow M(p)$ . Finally, these elements are structured by the syntactic complex  $q(p)$  into the formal configuration  $K(p) \subseteq M(p)$ , which is exactly described by the formal expression  $\varphi(p) = \text{fmlint}(u(p)) \in \text{Exp}(\text{XL})$ , where  $u(p) = q(p) \otimes \text{semrep}(q(p)) \in \text{Tns}(\text{NL})$  is the semantic tensor complex representing  $p$ . More precisely,  $K(p)$  is the minimal substructure of  $M(p)$  satisfying  $\varphi(p)$ , i.e.  $M(p) \supseteq K(p) \models \varphi(p)$ .

Again, in the simplest case conforming to MEAO semantics, the structure  $K(p)$  consists of

- an object  $A = \text{modref}_M(p)(v(p)_{\text{DMN}}) \in \text{Obj}(M(p))$  for a term  $v(p)_{\text{DMN}} \in \text{Trm}(\text{XL})$  of metatype  $\text{DMN} \in \text{Typ}(\text{PL})$
- an object  $B = \text{modref}_M(p)(v(p)_{\text{RNG}}) \in \text{Obj}(M(p))$  for a term  $v(p)_{\text{RNG}} \in \text{Trm}(\text{XL})$  of metatype  $\text{RNG} \in \text{Typ}(\text{PL})$
- a map  $f_B^A = \text{modref}_M(p)(v(p)_{\text{MAP}}) \in \text{Map}(M(p))$  for a term  $v(p)_{\text{MAP}} \in \text{Trm}(\text{XL})$  of metatype  $\text{MAP} \in \text{Typ}(\text{PL})$
- an element  $x_A = \text{modref}_M(p)(v(p)_{\text{OBJ}}) \in A$  for a term  $v(p)_{\text{ARG}} \in \text{Trm}(\text{XL})$  of metatype  $\text{ARG} \in \text{Typ}(\text{PL})$

e) an element  $x_B = f_B^A(x_A) \in B$

while the formal description  $\varphi(p)$  is simply the statement ' $x_B = f_B^A(x_A)$ '.

As the formal description of an abstract configuration  $K(p)$  in the internal model  $M(p)$ , the expression  $\varphi(p)$  is not an executable translate of the natural expression  $p$ ; instead, it is a formal interpretation of the precise conditions, relative to an abstract model of an operational environment, upon which an effective executable form of  $p$  may be constructed. The ultimate transition to an appropriate external operational environment is properly accomplished by means of the metasemantic protocol. In essence, the construction of a metaformal expression  $X(p) \in \text{Exp(PL)}$  as a fully effective translation of an expression  $p \in \text{Exp(NL)}$  is simply a machine interpretable codification of the satisfaction relation  $M(p) \models \varphi(p)$ .

Formally, the expression  $X(p) = \text{pclcod}(M(p), \varphi(p)) \in \text{Exp(PL)}$  is constructed by means of the *protocol encoding*  $\text{pclcod} : \text{Mod(XL)} \times \text{Exp(XL)} \rightarrow \text{Exp(PL)}$  based on the protocol interpretation  $\text{pclint} : \text{Sem(NL)} \rightarrow \text{Trm(PL)}$  introduced above. From this metaformal perspective, the internal model  $M(p) \in \text{Mod(XL)}$  appropriate for  $p$  is referenced by the term  $V(p)_{\text{ENV}} = \text{pclint}(c(p), \text{ENV}) \in \text{Trm(PL)}$ , and semantic objects  $x(p)_T = \text{semrep}(q(p)_T) \in \text{Sem(NL)}$ , which represent distinguished subterms  $q(p)_T = \text{pcltrm}(q(p), T) \in \text{Syn(NL)}$  of  $q(p) = \text{synrep}(p) \in \text{Syn(NL)}$  corresponding to *metatypes*  $T \in \text{Typ(PL)}$ , are interpreted as terms  $w(p)_T = \text{pclint}(x(p)_T) \in \text{Trm(PL)}$ . For the purpose of protocol consistency, it is significant to note that the formal terms from which  $\varphi(p) \in \text{Exp(XL)}$  is assembled are simply the interpretations  $v(p)_T = \text{fnltrm}(w(p)_T) \in \text{Trm(XL)}$ .

When PL is XMPL, the formal language associated with the universal protocol XMP, this metasyntactic construction is straightforward. Indeed, the syntax of XMPL is explicitly designed to accommodate MEAO semantics. In general, what is specified by any XMPL expression is an *operational environment*, a *transaction mapping*, a *mapping domain*, a *mapping range*, and a *mapping argument*. Refinements to any of these elements are indicated by appropriate nestings of subelements.

Again, in the simplest case conforming to MEAO semantics,  $X(p) \in \text{Exp(XMPL)}$  is the protocol statement

```
<XMP | <ENV | V(p) | ENV> <MAP | f(p) | MAP> <DMN | A(p) | DMN> <RNG | B(p) | RNG>
  <ARG | x(p) | ARG> | XMP>
```

where  $\langle ABC | xyz | ABC \rangle$  is a term of protocol type ABC and content xyz. The basic types occurring here are

XMP = transaction protocol  
 ENV = operational environment  
 MAP = transaction mapping  
 DMN = mapping domain  
 RNG = mapping range  
 ARG = mapping argument

For example, the XMPL statement which results from a metaformal interpretation of the sample natural language expression being processed is

```
<XMP | <ENV | email | ENV><MAP | send | MAP><DMN | <ADR | user@here.net | ADR> | DMN>
  <RNG | <ADR | bob@there.net | ADR> | RNG>
    <ARG | <MSG | <SBT | <LST | Appointment | LST> | SBT>
      <TXT | <LST | Are you going to go to your appointment by
        yourself? | LST> | TXT> | MSG> | ARG> | XMP>
```

using the additional protocol types

ADR = object address  
 MSG = message object  
 SBT = message subject  
 TXT = message text  
 LST = literal string

Note that the actual message text field of the XMPL statement consists of a question with the appropriate substitutions of 2<sup>nd</sup> person pronouns for the original 3<sup>rd</sup> person forms referring to the recipient "Bob". These transformations are computed as natural consequences of the syntactic relations coded into the algebraic form of the original natural language expression. In fact, all terms in the XMPL translation are computed similarly, i.e. as significant correlates of objects in the semantic complex structured by the syntactic dependencies indicated by the original natural language expression.

This metaformal syntactic scheme provides an effective universal template for abstract computations, and most significantly, for further translations into exact forms which are executable in specific operational environments.

*External Processing:* Refer to module 3.4 of FIG. 3.

*External Representation:* Refer to submodule 3.4.1 of FIG. 3.

Contact of the system with specific operational environments is made by means of an *external representation extrep* :  $Mod(XL) \rightarrow Env(NL)$  which associates appropriate external operational environments with internal formal models. In fact, these internal structures are explicitly designed as abstract models of external environments in order to accommodate this representation; accordingly, the external representation *extrep* may be viewed as the inverse of an internal representation *intrep* :  $Env(NL) \rightarrow Mod(XL)$  arising from a prior analysis of those operational environments which are relevant to the system.

For environments  $E \in Env(NL)$  which have their own executable language EL, this representation facilitates a translation between XL and EL by providing the semantic conditions which determine the mapping from terms of XL to the corresponding terms in EL. It is by means of this mediation in terms of its object language XL that the natural language NL, which is appropriate for informal negotiations in a wide variety of environments, becomes a metalanguage for an executable language EL, and therefore provides a basis for a meaningful operational semantics.

*External Interpretation:* Refer to submodule 3.4.2 of FIG. 3.

For a specific application in some external operational environment  $E \in \text{Env}(\text{NL})$  with an associated executable language  $EL$ , i.e. an application relative to a structure  $E$  for which there is an *executable interpretation*  $\text{expint}_E : \text{Exp}(EL) \rightarrow E$ , expressions of the object language  $XL$  are translated into expressions of  $EL$  by means of an *external translation*  $\text{exttrn}_E : \text{Exp}(XL) \rightarrow \text{Exp}(EL)$ . The commutative diagram shown in FIG. 7a illustrates the respective roles of the syntactic and semantic algebras for  $XL$  and the formal apparatus of  $XL$  in determining this translation.

The role of the metasemantic protocol indicated in this diagram is implicit. As an effective implementation of a formal interpretation scheme between  $XL$  and  $EL$ , it controls a critical aspect of the internal semantics of the system. As an effective translation medium between the internal formal structures of  $XL$  and the executable structures of  $EL$  associated with any operational environment  $E$  for  $XL$ , this protocol controls the external semantics of the system. The diagram shown in FIG. 7b illustrates the scope of these dual functions by highlighting the influence of the protocol on the system.

In particular, relative to an external operational environment  $E = \text{extrep}(M) \in \text{Env}(\text{NL})$  modeled by an internal structure  $M \in \text{Mod}(XL)$ , the formal translation  $\text{exttrn}_E : \text{Exp}(XL) \rightarrow \text{Exp}(EL)$  is implemented by means of the protocol encoding  $\text{pclcod} : \text{Mod}(XL) \times \text{Exp}(XL) \rightarrow \text{Exp}(EL)$  and translation  $\text{pcltrn} : \text{Exp}(EL) \rightarrow \text{Exp}(EL)$  as the composition  $\text{exttrn}_E(\varphi) = \text{pcltrn}(\text{pclcod}(M, \varphi))$  for any formal expression  $\varphi \in \text{Exp}(XL)$ ; moreover, the appropriate operational environment is determined by means of the protocol representation  $\text{pclrep} : \text{Exp}(EL) \rightarrow \text{Env}(\text{NL})$  as  $E = \text{pclrep}(\text{pclcod}(M, \varphi))$ . In short, all external transactions are mediated by the protocol.

Indeed, the universal transaction protocol XMP easily accommodates externalization since its associated formal language XMPL naturally translates into executable languages such as SQL (Standard Query Language) and SMTPL (the language of the mail protocol SMTP), and even executable extensions of XML (eXtensible Markup Language), in the manner indicated above, where  $EL$  is any of these executable languages. As such, XMPL forms a natural bridge between the internal semantics of the system and the external semantics of the environments in which it operates. The control structure of XMP then makes these external translations effective by finally facilitating in appropriate operational environments the execution of commands originally issued by the user in natural language.

More precisely, as given in the simple machine interpretable form

$\langle XMP \mid \langle ENV \mid V \mid ENV \rangle \langle MAP \mid f \mid MAP \rangle \langle DMN \mid A \mid DMN \rangle \langle RNG \mid B \mid RNG \rangle \langle ARG \mid x \mid ARG \rangle \mid XMP \rangle$

introduced above, a basic protocol statement  $X \in \text{Exp}(\text{XMPL})$  encodes the instruction to execute in environment  $E$  the operation  $f_E$  with domain  $A_E$  and range  $B_E$  on the argument  $x_E$ , where the external elements

- a)  $E = \text{extref}(X, ENV) = \text{pclrep}(X) \in \text{Env}(\text{NL})$
- b)  $f_E = \text{extref}(X, MAP) = \text{modref}_E(\text{extint}_E(f)) \in \text{Map}(E)$
- c)  $A_E = \text{extref}(X, DMN) = \text{modref}_E(\text{extint}_E(A)) \in \text{Obj}(E)$

d)  $B_E = extref(X, RNG) = modref_E(extint_E(B)) \in Obj(E)$   
 e)  $x_E = extref(X, ARG) = modref_E(extint_E(x)) \in A_E$

are properly interpreted by means of the uniform *external reference*  $extref : Exp(PL) \times Trm(PL) \rightarrow Env^*(NL)$ , or alternatively, initialized by the *protocol representation*  $pclrep : Exp(PL) \rightarrow Env(NL)$ , and then determined locally by the composition of the *model reference*  $modref_E : Trm(EL) \rightarrow E$  with the *external term interpretation*  $extint_E : Trm(PL) \rightarrow Trm(EL)$ , both of which are defined relative to the operational environment  $E \in Env(NL)$ . The actual external processing of this instruction is finally accomplished by application of the *execution process*  $envexc_E : Exp(EL) \rightarrow E$  to the formal translate  $\xi = pclrn(X) \in Exp(EL)$ , which is the executable expression encoded in  $X$  and constructed over the terms  $extint_E(f), extint_E(A), extint_E(B), extint_E(x) \in Trm(EL)$ . The result is a new state  $envexc_E(\xi) \in E$ .

For example, when interpreted according to its explicit formal specification as an email instruction to be executed under the control of the external protocol SMTP, the XMPL statement resulting from the sample expression being processed yields the result

To: Bob  
 From: <user>  
 Subject: Appointment  
 Text: Are you going to go to your appointment by yourself?

as a properly formatted message in the file `bob@there.net`, where `<user>` is the local (recipient's) identification of the sender (user).

#### APPLICATIONS AND TECHNOLOGY:

METASCIPT is currently implemented as a translation from a natural language (English) into the formal language (XMPL) associated with a universal transaction protocol (XMP: eXternal Media Protocol). In turn, this formal language is suitable for interpretation by digital components in external operational environments into executable machine instructions. Thus METASCIPT allows a human user to communicate naturally in an effective manner with (and through) any programmable device, hence networked configurations of such devices, compatible with the protocol XMP. In this capacity, METASCIPT is a *natural language interface* to any sufficiently capable digital environment, whether it be a single device such as a computer, a cellular phone, a PDA (Personal Digital Assistant), a kitchen appliance, an automobile, or a whole network of such devices such as a local intranet or the global Internet. As a complete NLP system, the combined technologies of METASCIPT and XMP enable a seamless integration of all participants, human and digital alike, into an effective ubiquitous network.

The fundamental algorithm upon which METASCIPT is based employs a *reduction* to formal syntactic structures over terms defined in an extensible *lexicon*. This term reduction incorporates both syntactic type and semantic context to achieve an effective formal representation and interpretation of the meaning conveyed by any natural language expression. Extensibility of the lexicon under specific user direction provides the capacity for the system to expand its knowledge of vocabulary and usage, and consequently, offers an effective mechanism under user control for establishing definite incremental

enhancements to the system's linguistic capabilities, hence substantially increasing the system's familiarity with (and competence in) particular operational environments.

In addition, the system automatically gains functional complexity through its object-oriented semantics, whereby the addition of formal terms having composite objects generated by algebraic representations of natural linguistic terms as direct references permits unlimited efficiency and sophistication of machine comprehensible natural language usage. Put simply, the system learns as it goes. Moreover, any desired level of syntactic disambiguation is attainable by increasing the local dimensionality of the underlying reduction matrix, though this feature is part of the underlying algorithm, and therefore independent of user modulation.

Finally, while the invention has been particularly shown and described with reference to preferred embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the spirit and scope of the invention.

**CLAIMS:**

- 1) A natural language processing apparatus for translating natural language into a formal language executable on a programmable device, said system comprising,
  - a) memory for storing data;
  - b) a data processor;
  - c) an input device for presenting natural language text to said system;
  - d) a text parser for partitioning said text into a sequence of sequences of string of characters or pretokens;
  - e) a lexicon for storing lexical terms as token associated with lexical type and reference data;
  - f) a lexical type assignment process for assigning lexical types to pretokens by comparison to terms in the lexicon;
  - g) a lexical insertion processor for inserting terms into the lexicon under specific control;
  - h) a control processor for invoking lexical insertions under the condition that a pretoken is not recognized as a lexical token;
  - i) a type contextualization processor by which refined lexical types may be reassigned to tokens depending on syntactic context;
  - j) a type reduction matrix;
  - k) a term reduction processor which uses said type reduction matrix to determine proper syntactic dependencies between tokens in a sentence;
  - l) a term inversion processor for constructing chains of syntactic dependencies among lexical terms in an expression and for determining the proper dependencies between those chains;
  - m) a syntactic tree generation processor for constructing syntactic trees representing the syntactic structure of each processed expression;
  - n) a syntactic algebra comprising syntactic terms formally representing processed expressions;
  - o) a syntactic representation processor for constructing syntactic terms to represent the formal syntactic structure of processed expressions;
  - p) a semantic algebra comprising semantic objects as formal references of appropriate terms in the syntactic algebra;
  - q) a semantic representation processor for associating internal semantic object references with terms in the syntactic algebra;

- r) a semantic tensor algebra comprising correlated pairs of syntactic algebraic terms and their semantic object representations;
- s) a formal representation processor for associating appropriate internal formal models with terms in the semantic tensor algebra;
- t) a formal interpretation processor for transforming terms in the syntactic algebra into equivalent expressions in an internal formal language;
- u) an external representation processor for associating external operational environments with internal formal models;
- v) an external interpretation processor for translating expressions in an internal in an internal formal language into equivalent formal expressions executable into appropriate external operational environments.

2) A method for translating natural language into a formal language executable on a programmable device, said method comprising the steps of:

- a) receiving natural language text;
- b) parsing said text into a sequence of sequences of pretokens;
- c) recognizing pretokens as tokens in the lexicon;
- d) inserting new terms into the lexicon under specific control;
- e) assigning types to pretokens to form lexical terms for further syntactic processing;
- f) reassigning lexical types to tokens based on syntactic context;
- g) correlating terms occurring in a set of expressions in order to replace indirect references by appropriate direct references;
- h) establishing syntactic dependencies between terms in an expression through a process of term reduction;
- i) constructing chains of syntactic dependencies and determining dependencies between those chains, by a process of term inversion;
- j) generating syntactic trees which represent the syntactic structures of said processed expressions;
- k) representing said processed expressions as terms in a syntactic algebra;
- l) representing terms in the syntactic algebra as objects in the semantic algebra;
- m) combining objects in the semantic algebra by means of a semantic product on pairs of semantic objects to form more complex semantic objects;

- n) representing correlated syntactic algebraic terms and semantic objects as terms in a semantic tensor algebra;
- o) representing terms in the semantic tensor algebra as internal formal models;
- p) transforming terms in the syntactic algebra into equivalent expressions in an internal formal language;
- q) associating external operation environments with internal formal models; and
- r) translating expressions of the internal formal language into equivalent formal expressions executable in an external operational environment.

3) In a natural language processing apparatus for translating natural language into a formal language executable on a programmable device, wherein said system includes processing means; input means for presenting natural language text to said system; a lexicon of terms; a text parser which partitions expressions into sequences of sequences of pretokens; a type assignment process for assigning syntactic types to pretokens by comparison to lexical terms in the lexicon and determining their status as tokens; a type contextualization process for reassigning lexical types to tokens based on syntactic context, a term correlation process for correlating terms occurring in a set of expressions in order to replace indirect references by direct references, said system comprising

- a) a type reduction matrix;
- b) a term reduction processor that uses the type reduction matrix to determine proper syntactic dependencies between tokens in an expression;
- c) a term inversion processor for constructing chains of syntactic dependencies among lexical terms in an expression and for determining the proper dependencies between those chains;
- d) a syntactic tree generation processor for constructing syntactic trees representing the syntactic structures of expressions;
- e) a syntactic algebra comprising syntactic terms formally representing processed expressions;
- f) a syntactic representation processor for constructing syntactic algebraic terms representing processed expressions;
- g) a semantic object algebra comprising semantic objects as internal references of terms in the syntactic algebra;
- h) a semantic product processor by which objects in the semantic object algebra are combined to form more complex semantic objects;
- i) a semantic representation processor by which internal semantic algebraic objects representing terms in the syntactic algebra are constructed;

- j) a semantic tensor algebra comprising correlated syntactic terms and semantic objects;
- k) a formal representation processor by which internal formal models are associated with terms in the semantic tensor algebra;
- l) a formal interpretation processor by which syntactic algebraic terms are transformed into equivalent expressions in an internal formal language;
- m) a semantic product processor by which objects in the semantic algebra are combined to form more complex semantic objects;
- n) an external representation processor by which external operational environments are associated with internal formal models; and
- o) an external interpretation processor by which expressions in an internal formal language are translated into equivalent formal expressions executable in an external environment;

4) A software system for translating natural language into a formal language executable on a programmable device, wherein said system includes processing means; input means for presenting natural language text to said system; a lexicon of terms; a text parser which partitions natural language texts into sequences of sequences of pretokens; a type assignment process for assigning syntactic types to pretokens by comparison to lexical terms in the lexicon and determining their status as tokens; a type contextualization process for reassigning lexical types to tokens based on syntactic context; a term correlation process for correlating terms occurring in a set of expressions in order to replace indirect references by direct references,

- a) a type reduction matrix;
- b) a term reduction process which uses the type reduction matrix to determine proper syntactic dependencies between tokens in an expression;
- c) a term inversion process for constructing chains of syntactic dependencies among lexical terms in an expression and for determining the proper dependencies between those chains;
- d) a syntactic tree generation process by which syntactic trees representing the syntactic structures of expressions are constructed;
- e) a syntactic algebra comprising syntactic terms formally representing processed expressions;
- f) a syntactic representation process by which syntactic algebraic terms representing processed expressions are constructed;
- g) a semantic object algebra comprising semantic objects as internal references of term in the syntactic algebra;
- h) a semantic object algebra comprising semantic objects as formal references of terms in the syntactic algebra;

- i) a semantic representation process by which internal semantic algebraic objects representing appropriate terms in the syntactic algebra are constructed;
- j) a semantic product process by which objects in the semantic algebra are combined to form more complex semantic objects;
- k) a formal representation process by which internal formal models object references are associated with terms in the semantic tensor algebra;
- j) a formal interpretation process by which syntactic algebraic terms are transformed into equivalent expressions in an internal formal language;
- l) an external representation process by which appropriate external operation environments are associated with internal formal models; and
- l) an external interpretation process by which expressions in an internal formal language are translated into equivalent formal expressions executable in an external operational environment.

5) A software system for a data processing device used in translating natural language into executable expressions in a formal language, wherein said data processing device includes a data processor and memory; input means for presenting natural language text to said system; a lexicon of terms; a text parser which partitions natural language texts into sequences of sequences of pretokens; a type assignment processor for assigning syntactic types to pretokens by comparison to lexical terms in the lexicon and determining their status as tokens; a type contextualization processor for reassigning lexical types to tokens based on syntactic context; a term correlation processor for correlating terms occurring in a set of expressions in order to replace indirect references by direct references; said software system comprising,

- a) a type reduction matrix for processing said expressions;
- b) a term reduction processor that uses the type reduction matrix to determine proper syntactic dependencies between tokens in an expression;
- c) a term inversion processor for constructing chains of syntactic dependencies among lexical terms in an expression and for determining the proper dependencies between those chains;
- d) a syntactic tree generation processor by which syntactic trees representing the syntactic structures of expressions are constructed;
- e) a syntactic algebra comprising syntactic terms formally representing said processed expressions;
- f) a syntactic representation processor by means of which syntactic algebraic terms representing processed expressions are constructed;
- g) a semantic object algebra comprising semantic objects as internal references of terms in the syntactic algebra;
- h) a semantic representation processor by which internal semantic algebraic objects representing terms

in the syntactic algebra are constructed;

- i) a semantic product processor by which objects in the semantic algebra are combined to form more complex semantic objects;
- j) a formal representation processor by which internal formal models are associated with terms in the semantic tensor algebra;
- k) a formal interpretation processor by which syntactic algebraic terms are transformed into equivalent expressions in an internal formal language;
- l) an external representation processor by which external operational environments are associated with internal formal models; and
- m) an external interpretation processor by which expressions in an internal formal language are translated into equivalent formal expressions executable in an external operational environment.

6) A system as in claim 5 further including a protocol by means of which

- a) selected ones of said internal formal models are associated with terms in said semantic tensor algebra;
- b) syntactic algebraic terms are transformed into equivalent expressions in the internal formal language;
- c) selected external operational environments are associated with selected formal models; and
- d) expressions in the internal formal language are translated into equivalent formal expressions executable in an external operational environment.

7) A system as in claim 5 further comprising

- a) a lexical insertion processor for inserting lexical terms into the lexicon under user control whereby said lexicon can be expanded and refined; and
- b) a controller for invoking lexical insertions under the condition that a pretoken is not recognized as a lexical token.

8) A system as in claim 7 further including a process control for inserting external lexical information to said lexicon to enable the system to learn new lexical information including vocabulary and associated lexical type and reference relations.

9) A data processing system for translating a natural language into a language executable as a formal machine language comprising, in combination,

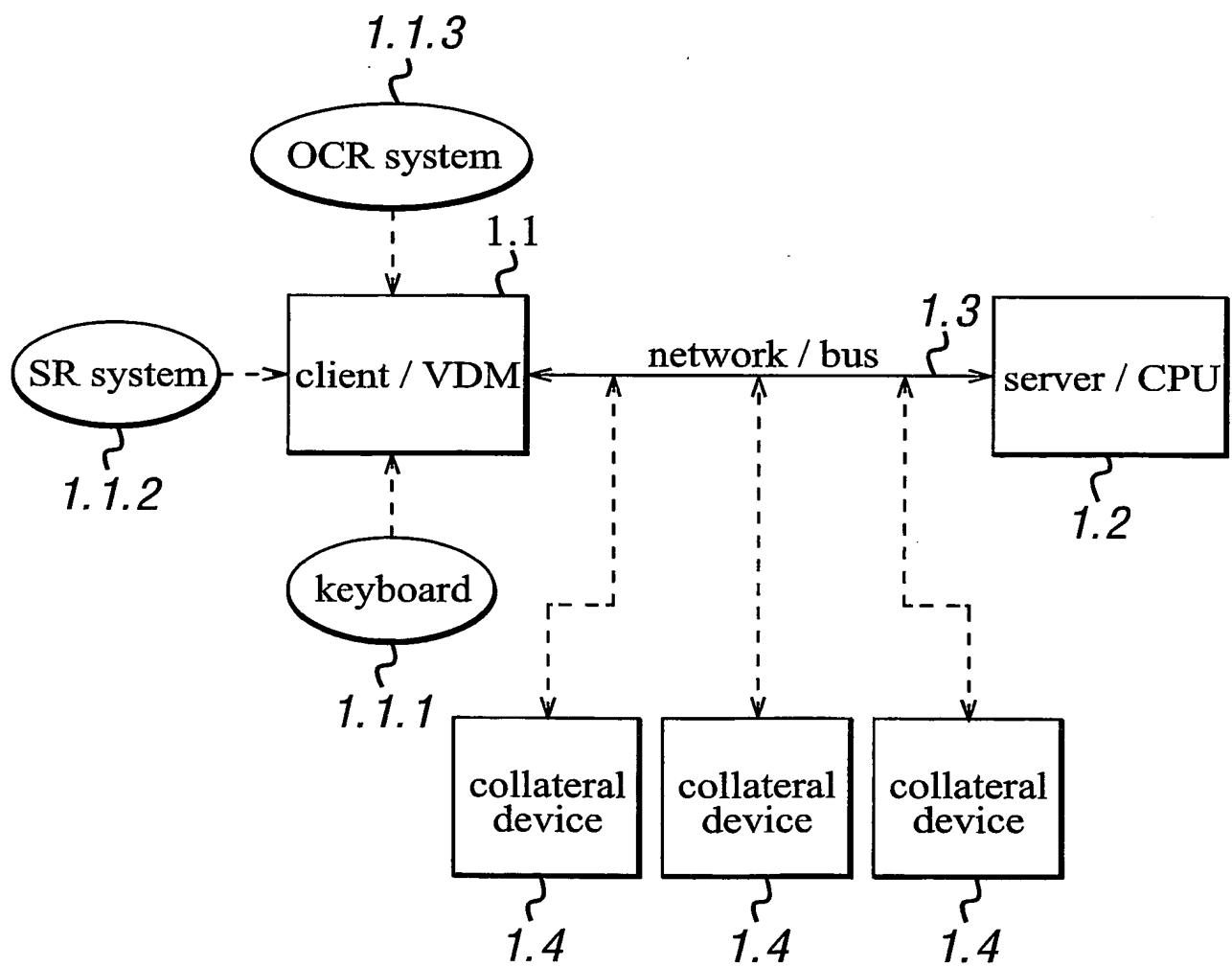
- a) input devices for inputting a natural language text to said system;
- b) text processing components for providing an output comprising a sequence of preexpressions based on said text;
- c) a syntactic processing component receiving said preexpressions and providing a sequence of syntactic complexes;
- d) semantic processing components for receiving said sequence of syntactic complexes and providing a sequences of formal expressions; and
- e) external processing components for providing a sequence of executable expressions to an external operational environment based on said formal expressions.

10. A method of translating a natural language into a language executable as a formal or machine language comprising the steps of,

- a) inputting a natural language text to a data processing system;
- b) providing an output comprising a sequence of preexpressions based on said text;
- c) receiving said preexpressions and providing a sequence of syntactic complexes;
- d) receiving said sequence of syntactic complexes and providing a sequences of formal expressions; and
- e) providing a sequence of executable expressions as an external operational structure based on siad formal expressions.

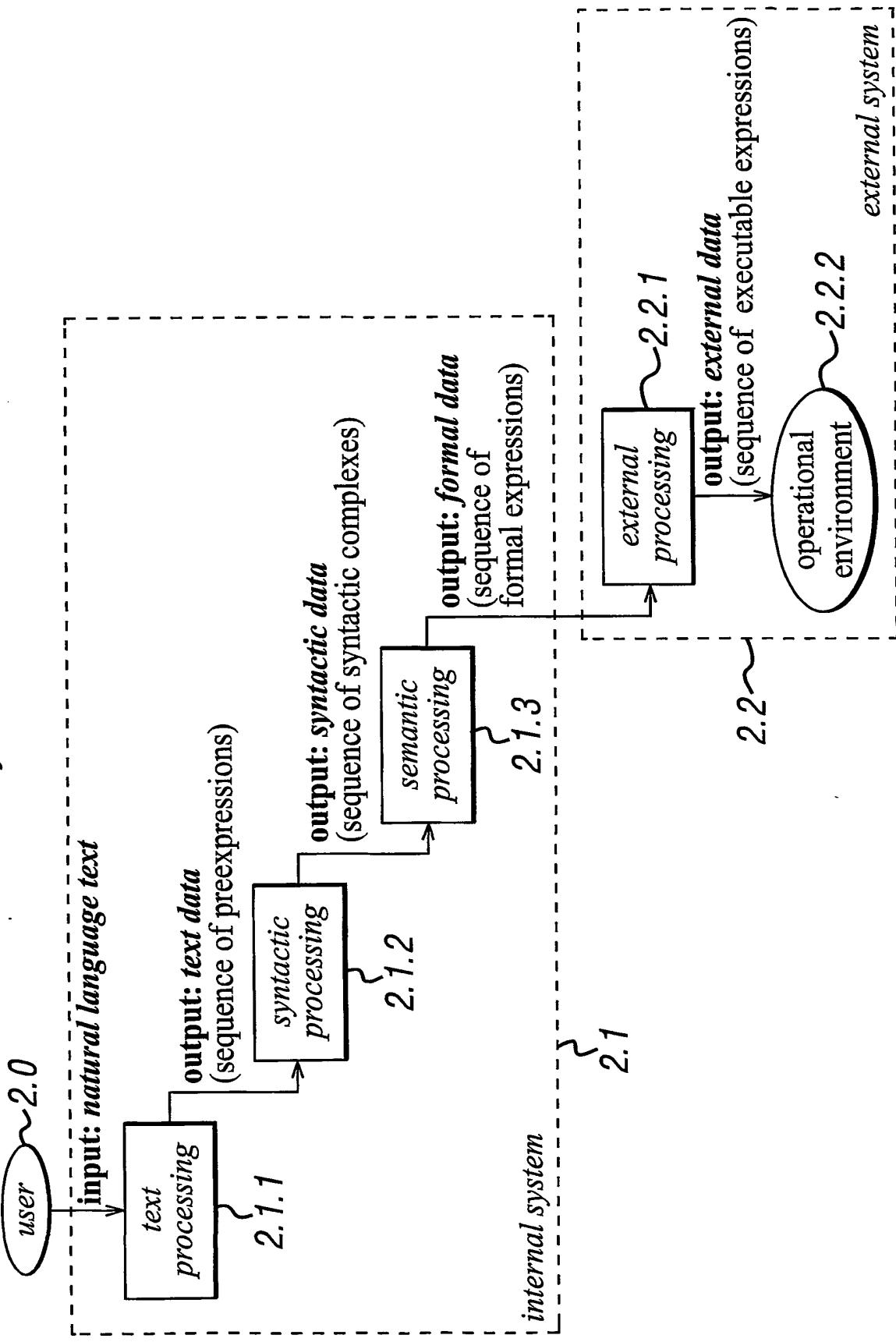
1/10

*Fig. 1*  
*Computer System Architecture*

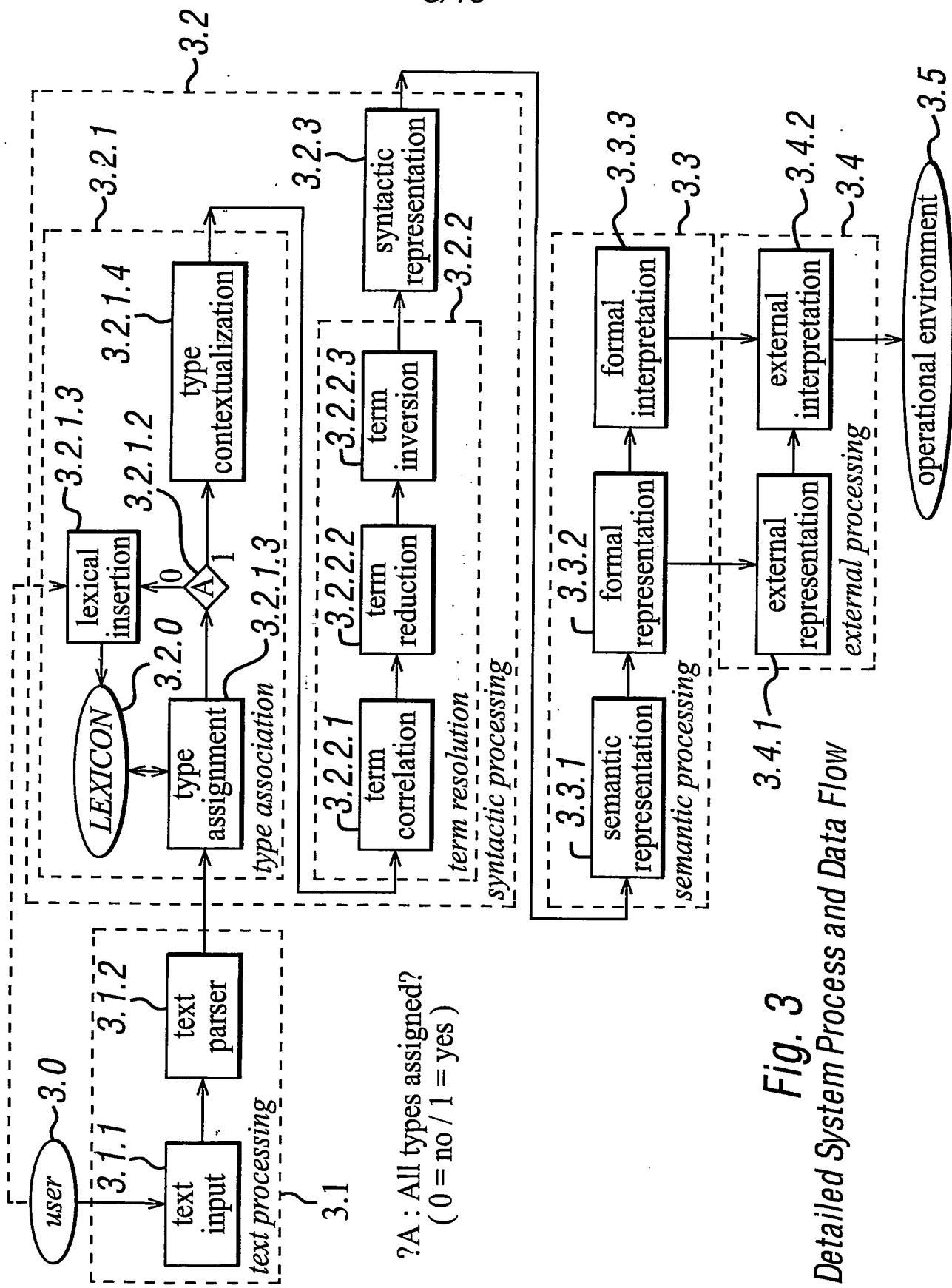


2/10

**Fig. 2**  
**General System Process and Data Flow**



3/10



**Fig. 3**  
*Detailed System Process and Data Flow*

4/10

*Fig. 4a Virtual Type Assignment*

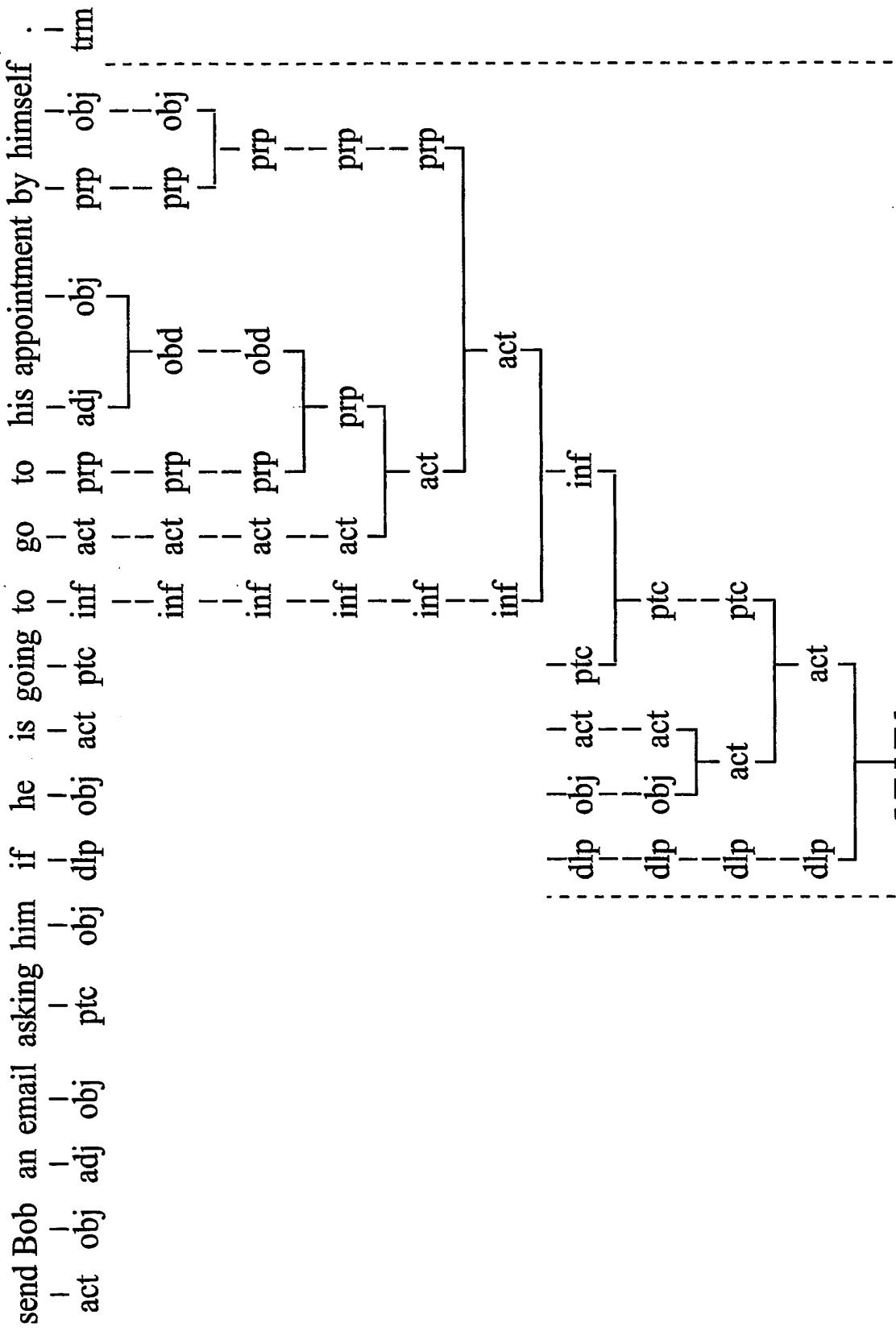
0) (send,act) = (send,lextyp(0,send)); act = action  
 1) (Bob,ppm) = (Bob,lextyp(0,Bob)); ppm = proper name/male  
 2) (an,adj) = (an,lextyp(0,an)); adj = adjective  
 3) (email,xao) = (email,lextyp(0,email)); xao = ambiguous action/object  
 4) (asking,ing) = (asking,lextyp(0,asking)); ing = ambiguous participle/gerund  
 5) (him,ppm) = (him,lextyp(0,him)); ppm = personal pronoun/male  
 6) (if,xdc) = (if,lextyp(0,if)); xdc = ambiguous delimiter/conditional  
 7) (he,ppm) = (he,lextyp(0,he)); ppm = personal pronoun/male  
 8) (is,sob) = (is,lextyp(0,is)); sob = state-of-being verb  
 9) (going,ing) = (going,lextyp(0,going)); ing = ambiguous participle/gerund  
 10) (to,xpi) = (to,lextyp(0,to)); xpi = ambiguous preposition/infinitive  
 11) (go,act) = (go,lextyp(0,go)); act = action  
 12) (to,xpi) = (to,lextyp(0,to)); xpi = ambiguous preposition/infinitive  
 13) (his,psm) = (his,lextyp(0,his)); psm = personal possessive/male  
 14) (appointment,xom) = (appointment,lextyp(0,appointment)); xom = ambiguous object/modifier  
 15) (by,prp) = (by,lextyp(0,by)); prp = preposition  
 16) (himself,prm) = (himself,lextyp(0,himself)); prm = personal reflexive/male  
 17) (.,trm) = (.,lextyp(0,.)); trm = termination

Fig. 4b Actual Type Assignment

0) (send,act) = (send,lextyp(0,send)); act = action  
 1) (Bob,pnm) = (Bob,lextyp(0,Bob)); pnm = proper name/male  
 2) (an,adj) = (an,lextyp(0,an)); adj = adjective  
 3) (email,obj) = (email,lextyp(1,email)); obj = object  
 4) (asking,ptc) = (asking,lextyp(1,asking)); ptc = participle  
 5) (him,ppm) = (him,lextyp(0,him)); ppm = personal pronoun/male  
 6) (if,dlp) = (if,lextyp(0,if)); dlp = phrase delimiter  
 7) (he,ppm) = (he,lextyp(0,he)); ppm = personal pronoun/male  
 8) (is,sob) = (is,lextyp(0,is)); sob = state-of-being verb  
 9) (going,ptc) = (going,lextyp(1,going)); ptc = participle  
 10) (to,inf) = (to,lextyp(2,to)); inf = infinitive  
 11) (go,act) = (go,lextyp(0,go)); act = action  
 12) (to,prp) = (to,lextyp(1,to)); prp = preposition  
 13) (his,psm) = (his,lextyp(0,his)); psm = personal possessive/male  
 14) (appointment,obj) = (appointment,lextyp(1,appointment)); obj = object  
 15) (by,prp) = (by,lextyp(0,by)); prp = preposition  
 16) (himself,prm) = (himself,lextyp(0,himself)); prm = personal reflexive/male  
 17) (.,trm) = (.,lextyp(0,.)); trm = termination

6/10

Fig. 5 Term Reduction Sequence



**SUBSTITUTE SHEET (RULE 26)**

7/10

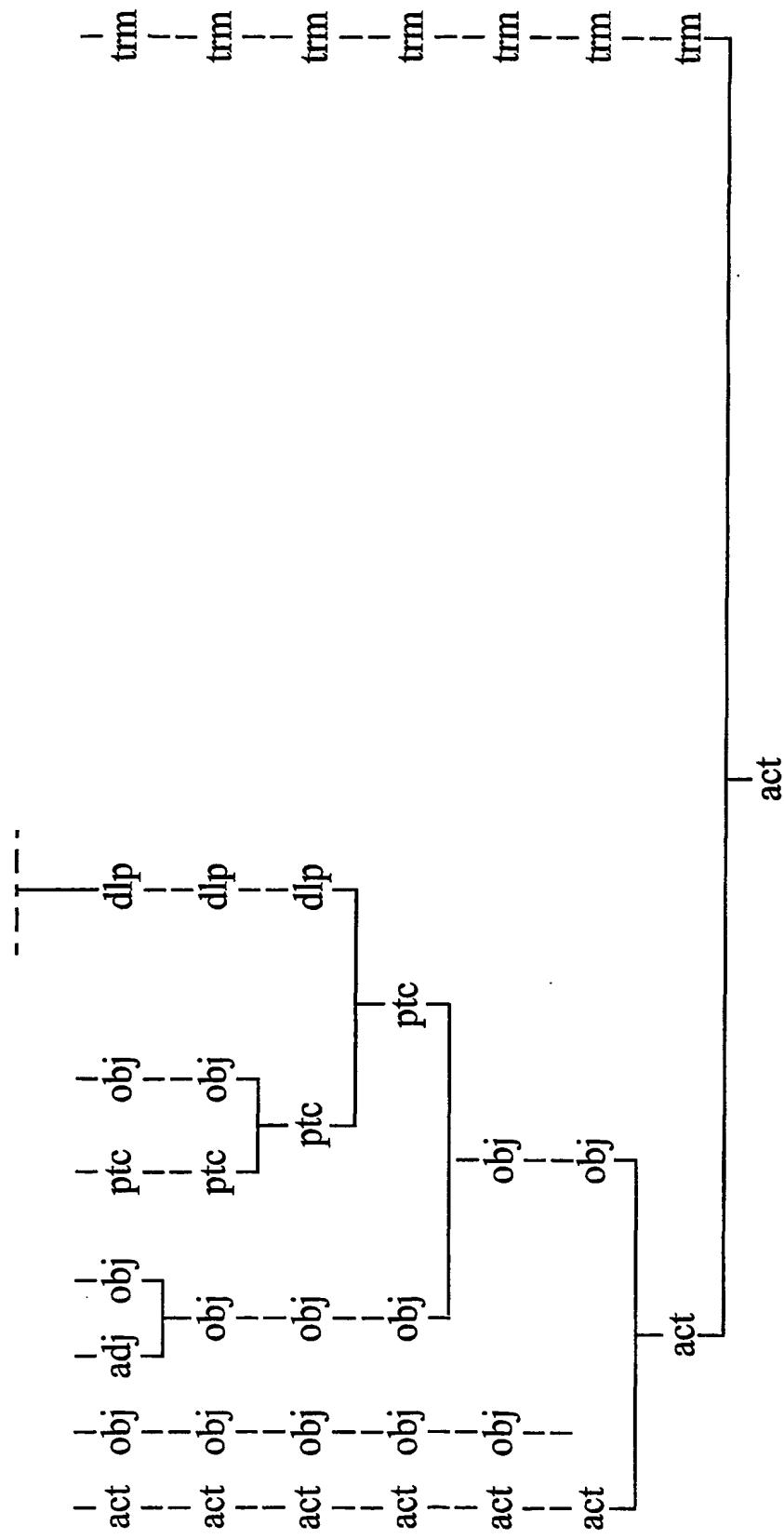


Fig. 5 (Contd.)

**SUBSTITUTE SHEET (RULE 26)**

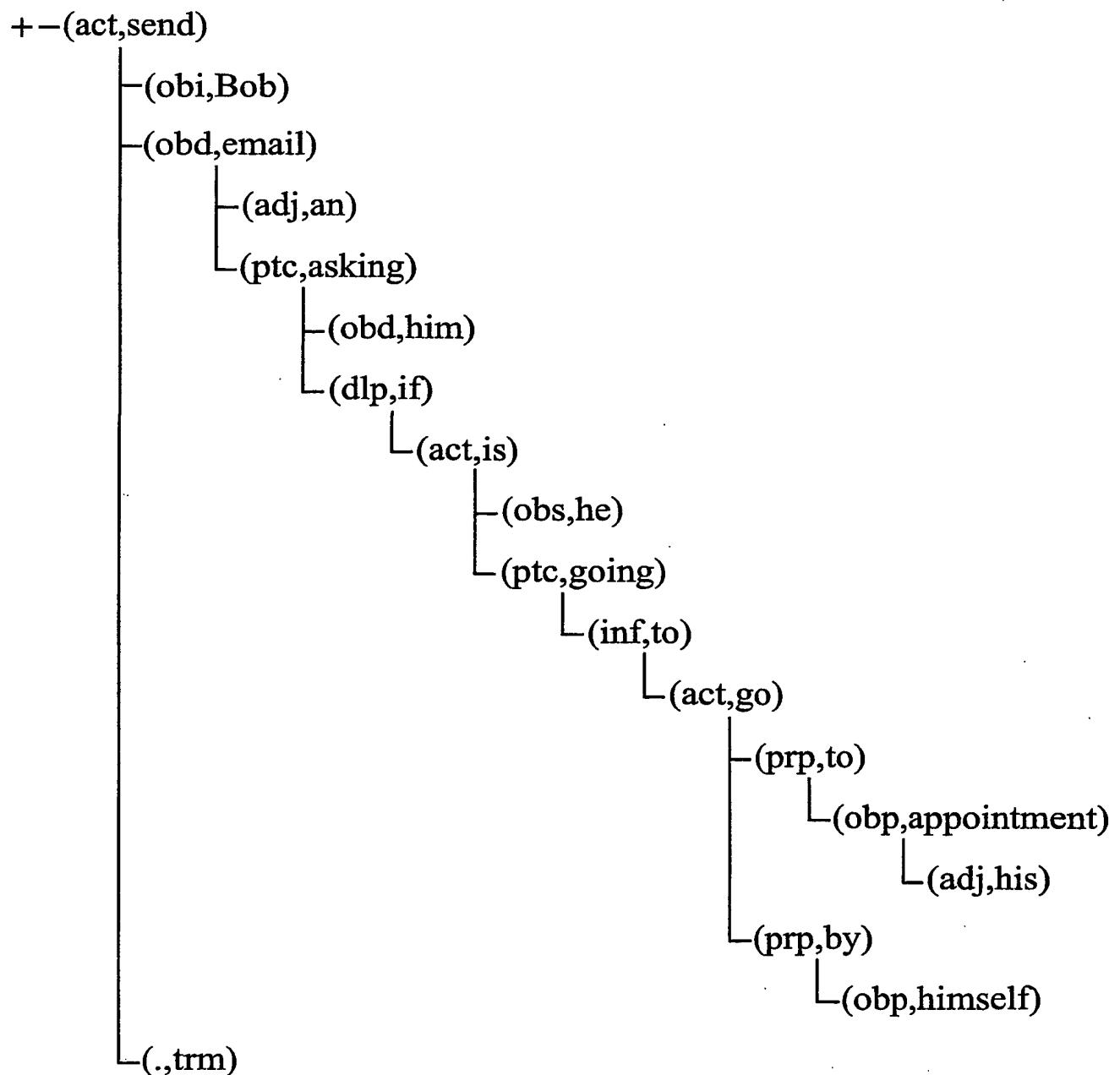
8/10

Fig. 6a Syntactic Dependency Chains

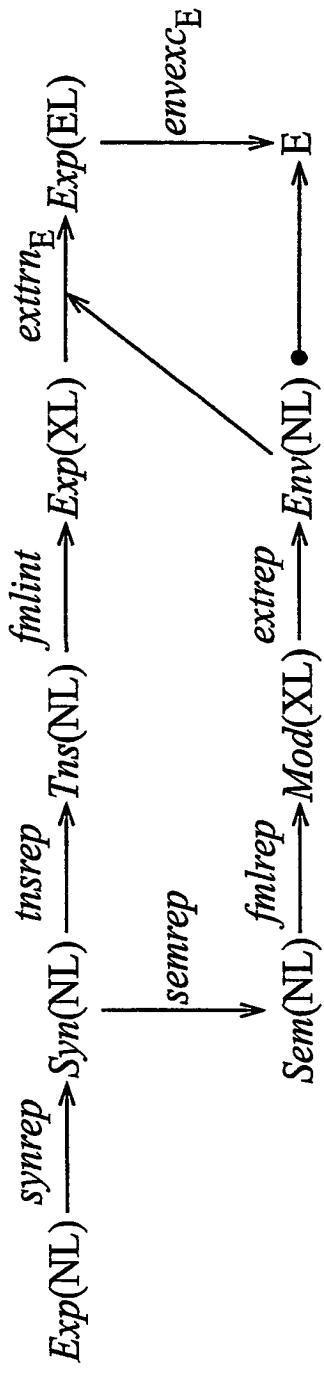
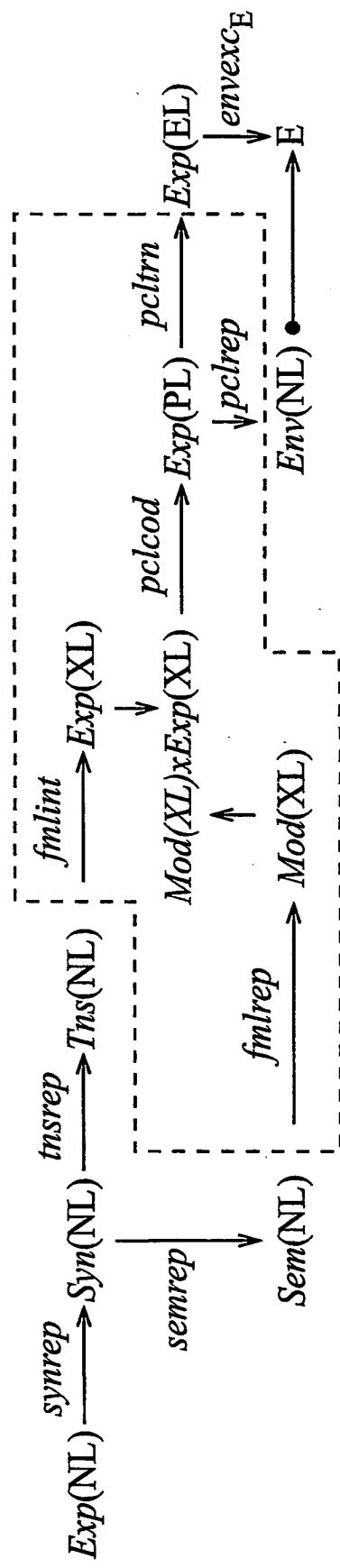
- 0) (send,act) → (Bob,obi)
- 1) (send,act) → (email,obd) → (an,adj)
- 2) (send,act) → (email,obd) → (asking,ptc) → (him,obd)
- 3) (send,act) → (email,obd) → (asking,ptc) → (if,dlp) → (is,act) → (he,obs)
- 4) (send,act) → (email,obd) → (asking,ptc) → (if,dlp) → (is,act) → (going,ptc) → (to,inf) → (go,act) → (to,ppr) → (appointment,obp) → (his,adj)
- 5) (send,act) → (email,obd) → (asking,ptc) → (if,dlp) → (is,act) → (going,ptc) → (to,inf) → (go,act) → (by,ppr) → (himself,obp)

SUBSTITUTE SHEET (RULE 26)

9/10

*Fig. 6b Syntactic Tree*

10/10

*Fig. 7a External Representation / Interpretation Schema**Fig. 7b Protocol Representation / Interpretation Schema*

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
27 December 2002 (27.12.2002)

PCT

(10) International Publication Number  
WO 02/103555 A3

(51) International Patent Classification<sup>7</sup>: G06F 17/27 (81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(21) International Application Number: PCT/GB02/02742 (84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

(22) International Filing Date: 12 June 2002 (12.06.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data: 09/883,693 18 June 2001 (18.06.2001) US

(71) Applicant (for all designated States except US): PIPEDREAM METASYSTEMS, INC. [US/US]; 1288 Columbus Avenue #153, San Francisco, CA 94133 (US).

(71) Applicant (for MG only): HOWE, Steven [GB/GB]; 36 Longacre, Chelmsford, Essex CM1 3BJ (GB).

(72) Inventor; and

(75) Inventor/Applicant (for US only): MANSON, Keith, S. [US/US]; P.O. Box 8406, Berkeley, CA 94707 (US).

(74) Agent: HOWE, Steven; Lloyd Wise, Commonwealth House, 1-19 New Oxford Street, London WC1A 1LW (GB).

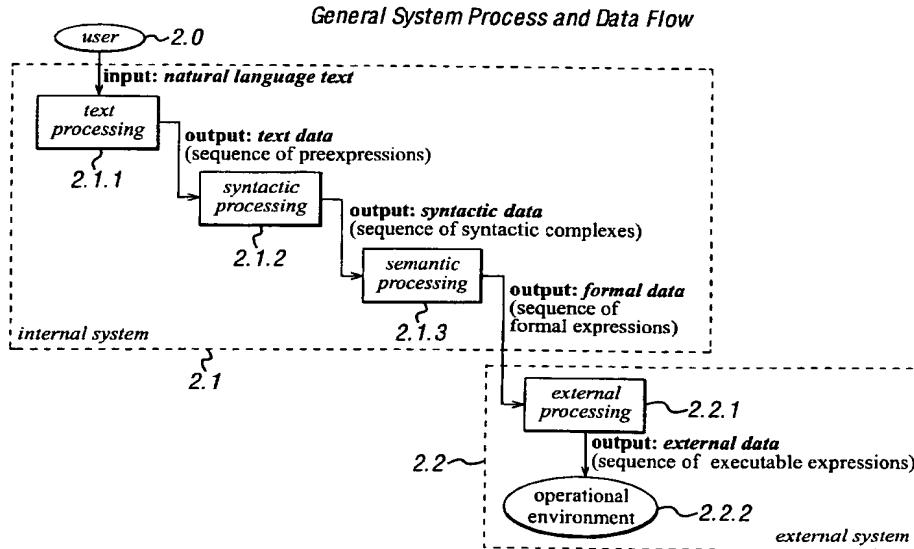
Published:

- with international search report
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments

(88) Date of publication of the international search report: 6 November 2003

*[Continued on next page]*

(54) Title: COMPUTER SYSTEM WITH NATURAL LANGUAGE TO MACHINE LANGUAGE TRANSLATOR



WO 02/103555 A3

(57) Abstract: Presented is a system and method for converting or translating expressions in a natural language such as English into machine executable expression in a formal language. This translation enables a transformation from the syntactic structures of a natural language into effective algebraic forms for further exact processing. The invention utilizes algorithms employing a reduction of sequences of terms defined over an extensible lexicon into formal syntactic and semantic structures. This term reduction incorporates both syntactic type and semantic context to achieve an effective formal representation and interpretation of the meaning conveyed by any natural language expression.



*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

## INTERNATIONAL SEARCH REPORT

International Application No  
PCT/GB 02/02742A. CLASSIFICATION OF SUBJECT MATTER  
IPC 7 G06F17/27

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)  
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	<p>RICHARDSON ET AL: "MindNet: Acquiring and Structuring Semantic Information from Text"</p> <p>ACL'98: 36TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS AND 17TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 'Online! vol. 2, May 1998 (1998-05), pages 1098-1102, XP002254590</p> <p>Retrieved from the Internet: &lt;URL: http://research.microsoft.com/research/pubs/view.aspx?tr_id=150&gt; 'retrieved on 2003-09-16! the whole document</p> <p>----</p> <p style="text-align: center;">-/-</p>	9,10

 Further documents are listed in the continuation of box C. Patent family members are listed in annex.

## \* Special categories of cited documents :

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier document but published on or after the international filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search

Date of mailing of the international search report

16 September 2003

29/09/2003

Name and mailing address of the ISA  
European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3016

Authorized officer

Triest, J

**INTERNATIONAL SEARCH REPORT**

International Application No  
PCT/GB 02/02742

**C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT**

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5 966 686 A (HEIDORN GEORGE ET AL) 12 October 1999 (1999-10-12) cited in the application abstract -----	1-5,9,10
A	US 6 246 977 B1 (DOLAN WILLIAM B ET AL) 12 June 2001 (2001-06-12) abstract -----	1-5,9,10

**INTERNATIONAL SEARCH REPORT**

Information on patent family members

International Application No

PCT/GB 02/02742

Patent document cited in search report	Publication date	Patent family member(s)		Publication date
US 5966686	A 12-10-1999	CN	1226327 A	18-08-1999
		DE	69710458 D1	21-03-2002
		DE	69710458 T2	27-06-2002
		EP	0907923 A1	14-04-1999
		JP	2000513473 T	10-10-2000
		WO	9800793 A1	08-01-1998
US 6246977	B1 12-06-2001	US	6076051 A	13-06-2000
		CN	1252876 T	10-05-2000
		EP	0965089 A1	22-12-1999
		JP	2001513243 T	28-08-2001
		WO	9839714 A1	11-09-1998
		US	6161084 A	12-12-2000

**THIS PAGE BLANK (USPTO)**